João Abel Rainho Fonseca

Licenciado em Biologia

Exploring the role of proteolysis in Extracellular Matrix remodeling: Links to Chronic Obstructive Pulmonary Disease and Lung Cancer

Dissertação para obtenção do Grau de Mestre em
Genética Molecular e Biomedicina

Orientador: Susana Seixas, PhD, Instituto de Investigação e Inovação em Saúde, Universidade do Porto (I3S); Instituto de Patologia e Imunologia Molecular, Universidade do Porto (IPATIMUP).

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Setembro 2017

Exploring the role of proteolysis in Extracellular Matrix remodeling: Links to Chronic Obstructive Pulmonary Disease and Lung Cancer

AKNOWLEDGMENTS

Começo por agradecer ao diretor do i3S, Professor Doutor Mário Barbosa, bem como ao diretor do IPATIMUP, Professor Doutor Manuel Sobrinho Simões, pela oportunidade de realizar o meu trabalho em ambos os institutos. Agradeço também à Doutora Luísa Pereira, por me ter recebido no seu grupo "Genetic Diversity".

À Doutora Susana Seixas, minha orientadora, pela oportunidade de me envolver este projeto, por todo o apoio, compreensão, disponibilidade, e tudo aquilo que me transmitiu ao longo deste ano, essencial para o desenvolvimento do projecto.

À Sílvia e à Patrícia, membros do grupo ao qual pertencia, pela disponibilidade para me ajudarem com qualquer dúvida que me surgisse durante o meu trabalho, pela simpatia e boa disposição, proporcionando sempre um bom ambiente no laboratório e um grande apoio ao longo deste ano.

À Joana, minha colega de mestrado, que em conjunto realizamos várias etapas do nosso trabalho, sendo que este apoio mútuo foi importante ao longo do ano. Ao meu compincha de laboratório, Alex, por todas as asneiras que fizemos juntos no laboratório, e de todas as vezes que nos apoiamos um ao outro para não desanimar.

Ao resto do grupo Genetic Diversity, por toda a ajuda prestada durante o desenvolvimento deste projecto, tendo sido essencial em vários pontos do percurso.

À minha família por todo o apoio ao longo do ano, incentivando-me a continuar a acreditar em mim mesmo quando tive dúvidas. Um grande obrigado mãe e pai, pela dedicação em fazer com que nunca fosse a baixo, a todos os meus tios (Luís, Lucília, Adelaide, Amélia, Francisco, Graça), ao meu avô Joaquim, à minha prima Catarina, à Mi e ao Gavina, à Sandra, por toda a preocupação a tentar saber "como está a correr essa coisa", sabem que o vosso apoio foi bastante importante. Um grande obrigado à minha irmã Inês, por me estar sempre a dizer que eu não faço nada, motivando-me assim a mostrar-lhe o contrário, e ao meu pequeno irmão André, por me proporcionar os momentos mais hilariantes deste ano, sendo a melhor escapatória da "chata vida de adulto". Por fim, obrigado "big bro" André, por me teres guiado sempre que precisei, considero-te um exemplo, e espero um dia chegar onde estás, ou pelo menos lá perto, desde que "seja aquilo que eu me veja a fazer".

Em último lugar, não podia deixar de agradecer a todos os meus grandes amigos que me acompanharam sempre ao longo destes anos e especialmente este último. Todas as jantaradas, farras, noitadas a trabalhar/estudar, ou pequenos momentos de simples contemplação de um momento, sabem que foram parte essencial da escapatória do stress deste ano.

ABSTRACT

Chronic obstructive pulmonary disease (COPD) and lung cancer (LC) are two complex disorders, currently representing the 4th cause of death and the most lethal cancer in Western countries, respectively. A mechanistic link between COPD and LC has been proposed due to an overlap of risk factors of both diseases, where uncontrolled proteolysis may be a critical event in their progress and outcomes. The activity of proteases, their substrates and inhibitors have a significant impact in the extracellular matrix (ECM) remodeling, which may ultimately lead to the development of COPD and LC. Despite the identification of several susceptibility factors in both diseases, there is still many aspects of their pathogenesis that require further elucidation. To address this issue, for our study, we selected 73 proteolysis genes, based on their roles in ECM remodeling, lung expression and/or presence in lung samples and former reports by Genome Wide Association Studies. In a first analysis, we took benefit of The Cancer Genome Atlas on-line database regarding clinical, epidemiological and mutational (somatic and germline) information for two common LC subtypes (adenocarcinoma and squamous cell carcinoma). We found that somatic mutability differs from germline trends and between the two LC subtypes, possibly affecting ECM in distinct ways. Then, we screened by means of PCR-based and Sanger sequencing techniques *SERPINB3/B4* and *CTSG* genes, in a small cohort of COPD and LC patients from which blood and bronchoalveolar lavage fluid samples were collected. Even though, we could not detect any somatic mutation in our sample, for *SERPINB3* we detect a considerable number of low-frequency variants in COPD cases in particular, suggesting a misfunction of this gene as a possible genetic risk factor for lung disease. Additional studies in larger cohorts of patients and controls are necessary to confirm this hypothesis.

Palavras Chave: Extracellular matrix; Proteolysis; Lung Cancer; Chronic Obstructive Pulmonary Disease; Inflammation; *SERPINB3*; *SERPINB4*; *CTSG*; *Genetic* variants.

RESUMO

A doença pulmonar obstrutiva crónica (DPOC) e o cancro do pulmão (CP) são duas doenças complexas, representando atualmente a quarta maior causa de morte e o cancro mais letal em países ocidentais, respetivamente. Tem sido sugerida uma associação mecanística entre a DPOC e o CP em parte devido à partilha de fatores de risco comuns em que a desregulação da proteólise pode também constituir um acontecimento crítico na sua evolução. A atividade das protéases, seus substratos e inibidores têm um impacto significativo na remodelação da matriz extracelular (MEC) o que em ultima análise pode levar ao desenvolvimento da DPOC e CP. Apesar de alguns fatores de suscetibilidade a ambas as doenças terem sido já reconhecidos, muitos aspetos da sua patogénese requerem um estudo mais aprofundado. Neste trabalho foram selecionados de 73 genes de proteólise, tendo em consideração o papel de cada um na remodelação da MEC, expressão ou presença em tecido pulmonar, e descrição por parte de estudos de associação genómicos. Numa primeira fase, foram extraídos da base de dados *The Cancer Genome Atlas* que compreende dois subtipos de CP (adenocarcinoma e carcinoma de células escamosas), informação clínica, epidemiológica e mutacional (somática e germinativa). Neste estudo verificou-se que a mutabilidade somática difere do padrão germinativo e entre os dois subtipos de CP, possivelmente afetando a MEC de forma distinta. Foi ainda efetuada uma análise dos genes *SERPINB3/B4* e *CTSG* por métodos de PCR e sequenciação de Sanger numa pequena coorte de doentes com DPOC e CP para os quais foram recolhidos sangue ou lavados brônquicos. Embora não tenha sido possível detetar qualquer mutação somática nas nossas amostras, para a *SERPINB3* foram detetadas diversas variantes de baixa frequência em casos de DPOC, sugerindo que alterações neste gene possam constituir um possível fator de risco genético na doença pulmonar. Estudos adicionais em coortes alargadas de doentes e controlos são essenciais para confirmar esta hipótese.

Palavras Chave: Matriz extracelular; Proteólise; Cancro do Pulmão; Doença Pulmonar Obstrutiva Crónica; Inflamação; *SERPINB3*; *SERPINB4*; *CTSG*; Variantes genéticas.

TABLE OF CONTENTS

ABBREVIATIONS xv

1. Introduction 1

1.1. State of the art of Human Genetics and Human Disease Studies 1

1.2. The extracellular Matrix (ECM) in the healthy lung 2

1.2.1. The ECM structural macromolecules: Elastin and fibrillar collagens 3

1.2.2. The ECM multiadhesive macromolecules: fibronectin and laminins 6

1.2.3. The role of proteolysis in the ECM remodeling 7

1.2.3.1. The Matrix metalloprotease (MMP) family 7

1.2.3.2. The families of Desintegrin and metalloproteases (ADAM and ADAMTS) 9

1.2.3.3. Cathepsins and other serine proteases 9

1.2.3.4. Protease Inhibitors 10

1.3. Complex Lung Diseases 12

1.3.1. Chronic Obstructive Pulmonary Disease 13

1.3.2. Lung Cancer 15

1.4. Mechanistic links between COPD and LC 16

1.4.1. Genetic Susceptibility factors 16

1.4.2. Oxidative stress, cell injury and inflammation 17

1.4.3. ECM remodeling and proteolysis in lung disorders 19

1.5. Patient Tailored Therapeutics 20

2. Materials and methods 23

2.1. Bioinformatics analysis 23

2.1.1. TCGA data – Lung Cancer 23

2.1.2. Clinical and epidemiological data analysis 23

2.1.3. Candidate genes selection 24

2.1.4. Variants expression and impact analysis 27

2.2. Screening of Portuguese COPD and LC cases 27
2.2.1. Samples 28
2.2.2. DNA Extraction 28
2.2.3. Polymerase Chain Reaction (PCR) amplification and sequencing 28
2.2.4. Sequence analysis 32
2.2.5. Statistical Analysis 32
3. Results and Discussion 33
3.1. TCGA data analysis 33
3.1.1.1. Epidemiological analysis 33
3.1.2. Proteolysis related genes analyses 38
3.1.2.1. Somatic and germline mutations rates 38
3.1.2.2. Candidate gene expression 44
3.2. Screening of *SERPINB3*, *SERPINB4* and *CTSG* genes in Portuguese COPD and LC cases . 48
3.3. Concluding remarks 52
4. References 53
5. Annexes 67

INDEX OF FIGURES

Figure 1.1. Conserved SERPIN three-dimensional structure 11
Figure 1.2. COPD Phenotypes 14
Figure 1.3. Repetitive cycles of tissue injury and repair. 19
Figure 2.1. Schematic representation of *SERPINB3/SERPINB4* amplification 30
Figure 3.1. Lung cancer incidence and mortality rates by geographical populations and gender (2012 data) 34
Figure 3.2. Mutation rates as retrieved by cBioPortal for ADC patients 38
Figure 3.3. Mutation rates as retrieved by cBioPortal for SCC patients 39
Figure 3.4. Top mutated candidate genes in ADC subtype and mutation functional predictions by Polyphen 40
Figure 3.5. Top mutated candidate genes in SCC subtype and mutation functional predictions by Polyphen 41
Figure 3.6. Top germline mutated candidate genes in ADC subtype and mutation functional predictions by Polyphen 43
Figure 3.6. Top germline mutated candidate genes in ADC subtype and mutation functional predictions by Polyphen 43
Figure 3.7. Top germline mutated candidate genes in SCC subtype and mutation functional predictions by Polyphen 43
Figure 3.8. Expression change of candidate genes in normal and tumor tissue, with normalization by a logarithmic scale of fold change 47
Figure 3.9. Schematic representation of *SERPINB3* 5'UTR variants location 49
Figure 3.10. *SERPINB3* protein structure with detected variants positions in reactive center loop highlighted 49
Figure A1. Somatic mutations rates of candidate genes in ADC patients with PolyPhen predictions .. 67
Figure A2. Somatic mutations rates of candidate genes in SCC patients with PolyPhen predictions ... 68
Figure A3. Germline mutations rates of candidate genes in ADC patients with PolyPhen predictions . 69
Figure A4. Germline mutations rates of candidate genes in SCC patients with PolyPhen predictions . 70
Figure A5. Somatic mutations rates of candidate genes in ADC patients with SIFT predictions 71
Figure A6. Somatic mutations rates of candidate genes in SCC patients with SIFT predictions 72
Figure A7. Germline mutations rates of candidate genes in ADC patients with SIFT predictions 73
Figure A8. Germline mutations rates of candidate genes in SCC patients with SIFT predictions 74
Figure A9. Somatic mutations rates of candidate genes in ADC patients with CADD predictions 75

Figure A10. Somatic mutations rates of candidate genes in SCC patients with CADD predictions 76
Figure A11. Germline mutations rates of candidate genes in ADC patients with CADD predictions .. 77
Figure A12. Germline mutations rates of candidate genes in SCC patients with CADD predictions ... 78

INDEX OF TABLES

Table 1.1. Summary of the human collagen classes..... 5
Table 1.2. Different MMPs clades and their corresponding ECM substrates..... 8
Table 1.3. Clade B SERRPINs with known roles in lung function 12
Table 2.1. Proteolysis related candidate genes selected for this study 24
Table 2.2. Primers used for the amplification of *SERPINB3/B4* genes..... 29
Table 2.3. Semi-nested PCR primers used for *SERPINB3/B4* amplification..... 30
Table 2.4. Primers used for SERPINB3/B4 and CTSGsequencing 31
Table 3.1. Distribution of lung cancer subtypes ADC and SCC per patient population ancestry..... 34
Table 3.2. Distribution of ADC and SCC cases per patient gender and ancestry..... 35
Table 3.3. Smoking history (PPY) and age of onset of ADC and SCC cases in each ancestry 36
Table 3.4. Tumor distribution per lung anatomic site for ADC and SCC cases..... 37
Table 3.5. COPD Stages of ADC and SCC cases, according to GOLD guidelines..... 37
Table 3.6. Significant tendency to co-occurrence of candidate genes in ADC cases 41
Table 3.7. Significant tendency to co-occurrence of candidate genes in SCC cases..... 42
Table 3.8. Variants identified in our cohort of Portuguese COPD and LC patients 50
Table 3.9. Statistical tests for low frequency variants found in *SERPINB3* and *SERPINB4* genes..... 51
Table T1. PCR conditions for SERPINB3/B4 amplification 79
Table T2. Semi-nested PCR conditions for SERPINB3/B4 amplification 79
Table T3. Sequencing PCR conditions for the three gene amplification..... 80

ABBREVIATIONS

AATD	Alpha-1 Antitrypsin Deficiency
ADAMs	a Disintegrin and Metalloprotease
ADAMTSs	a Disintegrin and Metalloprotease with Thrombospondin motifs
ADC	Adenocarcinoma
BALF	Bronchoalveolar lavage fluid
CNV	Copy number variant
COPD	Chronic Obstructive Pulmonary Disease
CTSG	Cathepsin G
CTSs	Cathepsin
DNA	Deoxyribonucleic Acid
ECM	Extracellular Matrix
EGF	Epidermal Growth Factor
ELANE	Neutrophil Elastase
ExAc	The Exossome Aggregation Consortium
FEV1	Forced Expiratory Volume
FVC	Forced Vital Capacity
GAG	Glycosaminoglycan
GOLD	Global Initiative for Chronic Obstructive Lung Disease
GWAS	Genome-wide Association Study
HGF	Hepatocyte Growth Factor
IBS	Iberian Population in Spain – 1000 Genomes
LC	Lung Cancer
MAF	Minor Allele Frequency
MMPs	Metalloprotease
NSCLC	Non-Small Cell Lung Cancer
PCR	Polymerase Chain Reaction
PG	Proteoglycans
PPY	Packs per Year
PRTN3	Proteinase 3

RCL	Reactive Center Loop
RNA	Ribonucleic Acid
ROS	Reactive Oxygen Species
SCC	Squamous Cell Carcinoma
SCLC	Small Cell Lung Cancer
SERPINA1	Alpha-1 Protease Inhibitor
SERPINB3	Squamous Cell Carcinoma Antigen 1
SERPINB4	Squamous Cell Carcinoma Antigen 2
SERPINs	Serine Protease Inhibitor
SNV	Single-nucleotide Variant
TCGA	The Cancer Genome Atlas
TGFα	Transforming growth factor alpha
TGFβ	Transforming growth factor beta
TIMPs	Tissue Inhibitor of Metalloprotease
UTR	Untranslated Region
WHO	World Health Organization

1. Introduction

1.1. State of the art of Human Genetics and Human Disease Studies

Understanding the molecular mechanisms of human disease and its genetic basis has been one of the main goals of the scientific community. In a global perspective, Mendelian disorders tend to be less prevalent in worldwide populations, more geographically confined and associated to single genes, where deleterious mutations arise in germline cells and are passed throughout generations rarely reaching polymorphic frequencies (MAF: minor allele frequency >1%). In contrast, complex (or multifactorial) disorders are in general the endpoint result of a combination of both genetic and environmental risk factors, and are often dispersed among diverse ethnic groups. Even though, the full extent of genetic variability associated to these common pathologies is not entirely acknowledged, this is more likely to be connected to germline mutations with a wide spectrum of frequencies (MAF<1% and MAF>1%) and variable contributions to disease susceptibility (small and large effect sizes) (Robinson et al. 2014; Mitchell 2012; Manolio et al. 2010).

Human cancers are by nature multifactorial disorders, however, another layer of complexity is added to these diseases, since a plethora of *de novo* mutations can originate in tumors (somatic mutations). These are usually classified into *driver* and *passenger* mutations, according to their outcomes in cancer progression. While *driver* mutations are accepted to confer a selective advantage, having critical roles in tumor growth, often reaching higher prevalence within tumor; *passengers* mutations are believed to behave neutrally not contributing to tumor clonal expansion (Merid et al. 2014; Stratton et al. 2009).

Over the last decades, the advent of several high-throughput genotyping and sequencing techniques, turned the study of human genetic variability (disease patients and controls) an easier task. In this field, several international consortiums released an unprecedented amount of data for major human groups (Africans, Europeans and Asians), starting by The HapMap Project (Frazer et al. 2007) in mid 2000s, which made available the information for millions of common (MAF>1%) single nucleotide polymorphisms (SNPs); and more recently by the 1000 Genomes Project that is providing a detailed overview of genome variants including less frequent single nucleotide substitutions, small insertion and deletions and large structural copy-number variants (CNVs). Moreover, the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP) and Exome Aggregation Consortium (ExAC) are also contributing to this data collection by surveying

mainly protein-coding regions at high depth for thousands of subjects screened in the scope of different disease-specific and population genetic studies (Lek et al. 2016; Auton et al. 2015). In addition , "The Cancer Genome Atlas" (TCGA) database, resulting also from a collaborative project between the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI) in the USA, is supporting the publication of genetic, clinical and other relevant data for 33 tumor types, comprising more than 11.000 patients (Auton et al. 2015; Chang et al. 2013).

Overall, these resources have already contributed to thousands of studies by consortiums themselves and by independent researchers, enabling a comprehensive analysis of the prevalence of many mutations associated to Mendelian and complex diseases. Nevertheless, the architecture of the complex diseases is not yet fully resolved and most variants identified so far seem to explain only a small fraction of heritability, even when using very large cohorts of patients and controls. Several hypotheses have been raised to explain this missing heritability in complex diseases, including the low power to detect gene-gene interactions; the inadequate accounting of shared environments between individuals, and the presence of a large number of variants with small impact in the disease onset, but also the occurrence of rare variants with stronger effects in subject health status (Manolio et al. 2010).

In this scope, this work is focused in the study of two multifactorial disorders affecting the respiratory system, Chronic Obstructive Pulmonary Disease (COPD) and lung cancer (LC), which are prime examples for the interaction between environmental and genetic factors, in disease susceptibility and pathogenesis (gene-by-environment theory). Here, we propose to address this issue by centering our variant screening in proteolysis related genes (proteases, their inhibitors and substrates) and their potential repercussions in the lung extracellular matrix (ECM). We will take advantage of published datasets and our own sample collection.

1.2. The extracellular Matrix (ECM) in the healthy lung

The extracellular matrix (ECM) is a bioactive milieu that provides structural support to cells and has important roles in regulating tissue morphogenesis, differentiation and homeostasis. ECM is mainly composed by water, polysaccharides, and proteins, including major structural proteins, such as elastin and collagens, and multiadhesive molecules, like fibronectin and laminins. However, the organization of ECM within and between tissues may vary due to the existence of different biochemical and biophysical dynamics and relationships between ECM cellular elements and the surrounding microenvironment (Mouw et al. 2015; Frantz et al. 2010). In this respect, alterations in

the ECM assemblage can have a large impact in tissue patterning during organogenesis and angiogenesis (Kniazeva & Putnam 2009; Mammoto et al. 2009) and, on the other hand, ECM proteins can influence cell behavior through signaling, binding to growth factors, mediating cell-adhesion and transducing signals into cells (Hynes 2013). One of the key features of the ECM is its ability to respond to diverse physiological and stress stimuli in a biological process often referred as ECM remodeling, in which damaged and proteolytic cleaved proteins are replaced by new ones (Swinehart & Badylak 2017; Bachman et al. 2015).

In the lung, the ECM is mainly responsible for the supportive scaffold of the alveolar wall, branching morphogenesis, and tissue repair after injury (Watson et al. 2016). Moreover, in the respiratory system, the ECM is also specialized for gas changes, the primary function of the lung, while also providing structural support to prevent airway collapse (Balestrini et al. 2016; Parameswaran et al. 2006). For this reason, one of the most important ECM components in the lung are elastin fibers, which confer stretch and recoil properties to pulmonary tissues; and collagen fibers, responsible for parenchyma support and basement membrane barrier functions (Dunsmore et al. 1996). Other key elements of the lungs are fibronectin and laminin fibrils, fundamental for cell-adhesion to the basement membrane and cell survival (Mouw et al. 2015) and glycosaminoglycans (GAG) that together with proteoglycans (PG) control cellular and macromolecule movements, water retention, ion content, and growth factor levels (Papakonstantinou & Karakiulakis 2009).

1.2.1. The ECM structural macromolecules: Elastin and fibrillar collagens

Elastin and fibrillar collagens are the main structural components of ECM.

In the lungs, in particular, the elastin provides natural stretching and contractile functions needed to respiratory cycles (Pelosi et al. 2007). This macromolecule, composed mainly by tropoelastin monomers (approximately 60-70 kDa), is encoded by *ELN*. As a biopolymer, elastin result from the aggregation of several monomers attached to each other by lysine residues. In the ECM elastogenesis, the presence of microfibrils congregating cysteine-rich proteins, such as fibrilin-1 and fibrilin-2, and microfibril-associated glycoprotein-1, is critical. Altogether, these macromolecules afford a scaffold for the deposition and assembly of tropoelastin into elastin fibers (Mithieux & Weiss 2006; Starcher et al. 1986) that in pulmonary tissues can be synthesized by different types of cells including chondroblasts, myofibroblasts and smooth muscle cells (Starcher

2000). Importantly, during adulthood, such elastogenesis processes are known to drop significantly leading to damage in elastic fibers to be irreversible and remain unrepaired (Humphrey et al. 2015).

In humans there are 28 known subtypes of collagen (Table 1.1). Generally, collagens are organized in homotrimers or restricted heterotrimers arranged in triple-helixes of α-chains. Briefly, these polypeptide chains are arranged in several repeat units of Gly-X-Y triplets (where X and Y denote any other amino acid residue) flanked by N- and a C- terminal propeptides. These ECM macromolecules can be further subdivided into fibrillar and non-fibrillar collagens (which can form or not fibril bundles, respectively), being the former the main collagen fibers found in lungs, more exactly collagen type I and III that provide pulmonary tissues, their tensile strength (Mouw et al. 2015; Rozario & Desimone 2011; Rocco et al. 2001). Other common types of collagen in the lung ECM are the non-fibrillar type IV collagen, present in alveolar and airways basement membranes, where these are thought to confer important barrier functions; and collagens types VII and XV that take part in anchoring of fibrils, linking basement membrane and connective tissue (Dunsmore 2008; Dunsmore et al. 1996). Along with their major structural properties collagens may also influence tissue development cell adhesion and migration (Hynes & Naba 2012; Frantz et al. 2010). Collagens ability to undergo molecular processes of turnover, including not only their synthesis, but also their deposition and degradation, is fundamental to the ECM dynamics of remodeling and thus, essential to the maintenance of healthy lungs (Humphrey et al. 2015; Pelosi et al. 2007).

Table 1.1. Summary of the human collagen classes.

Class	Collagen Type	Collagen-protein encoding genes
Fibrillar	I	*COL1A1*, *COL1A2*
	II	*COL2A1*
	III	*COL3A1*
	V	*COL5A1*, *COL5A2*, *COL5A3*
	XI	*COL11A1*, *COL11A2*, *COL11A3*
	XXIV	*COL24A1*
	XXVII	*COL27A1*
Fibril-associated collagens with interrupted triple helices (FACIT)	IX	*COL9A1*, *COL9A2*, *COL9A3*
	XII	*COL12A1*
	XIV	*COL14A1*
	XVI	*COL16A1*
	XIX	*COL19A1*
	XX	*COL20A1*
	XXI	*COL21A1*
	XXII	*COL22A1*
Basement membrane	IV	*COL4A1*, *COL4A2*, *COL4A3*, *COL4A4*, *COL4A5*, *COL4A6*
Long chain	VII	*COL7A1*
Filamentous	VI	*COL6A1*
Short chain	VIII	*COL8A1*
	X	*COL10A1*
Multiplexins	XV	*COL15A1*
	XVIII	*COL18A1*
Transmembrane domain (MACIT)	XIII	*COL13A1*
	XVII	*COL17A1*
	XXIII	*COL23A1*

In an histological overview of lung tissues, a high percentage of collagens is observed in large bronchi, small airways and large blood vessels like the pulmonary artery; whereas, elastin can be more usually detected in lung parenchyma, bronchi and blood vessels, as well (Balestrini et al. 2016; Townsley 2012; Parameswaran et al. 2006; Pierce & Hocott 1959).

1.2.2. The ECM multiadhesive macromolecules: fibronectin and laminins

Fibronectin and laminins are multidomain glycoproteins, whose major functions are to promote the adhesion between ECM structural components, across the later and soluble molecules in the extracellular space, as well as between cells and ECM. In other words, these multiadhesive molecules are capable of binding to other ECM proteins, cell surface receptors and growth factors through specific motifs found in their protein structure (Mouw et al. 2015).

Fibronectin is one of the most important ECM glycoproteins, which is involved in the interstitial organization of ECM and facilitates cell attachment, migration and differentiation (Schwarzbauer & Desimone 2011; Smith et al. 2007). Cellular fibronectin, the isoform most commonly found in the ECM, is secreted mostly by fibroblasts and organized in dimers of 250 kDa. In addition, cellular fibronectin is characterized by the presence of a 70 kDa N-terminal domain, responsible for fibril assembly and binding to the cell surface; and by V region, a key domain for cell motility and matrix assembly that also contains a α4β1 integrin binding site (To & Midwood 2011; Mao & Schwarzbauer 2005; Pankov & Kenneth 2002).

Laminins are ECM glycoproteins found essentially in basal membranes, intervening in the ECM-cell interactions, through the binding of cell surface receptors to ECM components (Lu et al. 2011; Rozario & Desimone 2011). Laminins are heterotrimeric proteins with up to 16 distinct isoforms, composed by five α-, four β- and three γ-chains subunits. Whereas α-chains are encoded by *LAMA1/2/3* genes, β-chains and γ-chains are expressed through *LAMB1/2/3* genes *LAMC1/2/3*, respectively. Similarly to fibronectin, laminins have large weights (400 to 900 kDa) and comprise as well binding domains to integrin and collagens, to maintain ECM-cell adhesion (Mouw et al. 2015; Hamill et al. 2009).

1.2.3. The role of proteolysis in the ECM remodeling

Multiple proteases and their inhibitors are essential in ECM remodeling to replace damaged macromolecules (protease substrates) and to maintain a fine balance between protein degradation and turnover. Proteases with key activities in ECM remodeling include different classes of metalloproteases, namely matrix metalloproteases (MMPs), desintegrin and metalloprotease domain containing proteins (ADAMs), and desintegrin and metalloprotease with thrombospondin motifs (ADAMTSs); several cysteine and serine proteases: such as neutrophil elastase (ELANE), cathepsin G (CTSG) and proteinase 3 (PRTN3). Whereas tissue inhibitor of metalloproteases (TIMPs) are able to control the activity of MMPs and ADAMs; the family of serine protease inhibitors (SERPINs) efficiently regulates diverse serine and cysteine proteases.

1.2.3.1. The Matrix metalloprotease (MMP) family

The MMP family comprises 23 members that are generally secreted as inactive zymogens and later activated by other MMPs or serine proteases, such as plasmin or neutrophil elastase (ELANE). Briefly, MMPs can be regulated at four different levels: by transcriptional and post-transcriptional regulation of gene expression; in the tissue milieu by proteolytic activation (removal of a propeptide segment); and by specific protease inhibitors (Löffek et al. 2011). During the normal remodeling (wound and healing) process, several MMPs are produced by neutrophils, macrophages, and wounded cells to degrade damaged ECM macromolecules – MMPs substrates.

Broadly speaking, MMPs can be allocated into six clades, according to their affinities towards different substrates, localization and structural organization (Table 1.2) (Caley et al. 2015; Mocchegiani et al. 2011). MMPs key features also include a N-terminal signaling anchor; a propeptide region, required for enzymatic activation; a calcium-dependent catalytic site composed by three histidines in a complex with a zinc ion, a hemopexin-like C-terminal domain, and a linker region which connects the catalytic and hemopexin-like C-terminal domains (Cathcart et al. 2015). Still, some members may lack some of these domains such as MMP-7, -23, and -26, which do not have linker and hemopexin domains. While other MMPs, such as MMP-9 and -21 can include additional functional domains (e.g. fibronectin- or vitronectin-like) domains (Caley et al. 2015; Cathcart et al. 2015).

Table 1.2. Different MMPs clades and their corresponding ECM substrates [Adapted from (Cathcart et al. 2015)].

Clade	MMP	ECM Substrates
Collagenases	MMP-1	Collagens (type I, II, III, VII, VIII, X, XI), fibronectin, laminin, vitronectin, entactin, gelatin, tenascin, aggrecan, and others
	MMP-8	Collagens (type I, II, III), aggrecan
	MMP-13	Collagens (type I, II, III, IV, VI, IX, X, XIV), fibronectin, gelatin, aggrecan, and others
	MMP-18	Collagen type I (rat)
Gelatinases	MMP-2	Collagens (I, II, III, IV, V, VII, X, XI), elastin, fibronectin, gelatin, laminin, vitronectin, tenascin, and others
	MMP-9	Collagens (IV, V, XI, XIV), elastin, laminin, vitronectin, and others
Stromelysins	MMP-3	Collagens (III, IV, V, VII, IX, X, XI), elastin, fibronectin, laminin, vitronectin, tenascin, and others
	MMP-10	Collagens (III, IV, V), elastin, fibronectin, aggrecan, and others
	MMP-11	Collagen type IV, fibronectin, laminin, gelatin
Matrilysins	MMP-7	Collagens (I, IV), elastin, fibronectin, vitronectin, laminin, gelatin, entacin, and others
	MMP-26	Gelatin, fibronectin, vitronectin
Membrane-type MMPs	MMP-14	Collagens (I, II, III), gelatin, fibronectin, tenascin, laminin, and others
	MMP-15	Fibronectin, tenascin, entacin, laminin, aggrecan, perlecan
	MMP-16	Collagen type III, gelatin, fibronectin, vitronectin, laminin
	MMP-24	Fibronectin, gelatin, chondroitin and dermatane sulphate proteoglycans
	MMP-17	Gelatin
	MMP-25	Collagen type IV, gelatin, fibronectin, chondroitin and dermatane sulphate proteoglycans
Others	MMP-12	Collagens (I, IV, V), elastin, gelatin, fibronectin, laminin, vitronectin, entactin, and others
	MMP-19	Collagen type IV, gelatin, laminin, entactin, and others
	MMP-20	Amelogenin, aggrecan
	MMP-21	Gelatin
	MMP-23	Gelatin
	MMP-27	Gelatin
	MMP-28	

1.2.3.2. The families of Desintegrin and metalloproteases (ADAM and ADAMTS)

The ADAM family comprises a total of 21 related molecules, in which only 13 show proteolytic activity, these comprise ADAM8-10, 12, 15, 17, 19-21, 28, 30 and 33 (Duffy et al. 2011). Structurally, ADAMs are organized in eight major units: the pre- and propeptide regions; and six other domains with functions as metalloprotease, disintegrin, cysteine-rich, Epidermal Growth Factor (EGF)-like, transmembrane, and cytoplasmic proteins (Giebeler & Zigrino 2016; Duffy et al. 2011). On the other hand, the ADAMTS family includes 19 members with some common features to ADAMs family, aside from the addition of innumerous thrombospondin motifs in the C-terminal domain, and the lack EGF-like, transmembrane and cytoplasmic domains (Kelwick et al. 2015; Paulissen et al. 2009). Based on their domain organization and known functions, ADAMTSs can be divided in 8 subgroups, being the most important the aggrecanase and proteoglycanase group (ADAMTS1, 4, 5, 8, 9, 15, 20), which cleave hyaluronic-binding chondroitin sulfate proteoglycan extracellular proteins. On the other hand, the group of pro-collagen N-peptidases (ADAMTS2, 3, 14) processes pro-collagen molecules (Kelwick et al. 2015). Similarly to MMPs the superfamily of ADAMs and ADAMTS also contain a catalytic domain bound to a zinc-ion (Kelwick et al. 2015; Tallant et al. 2010; Edwards et al. 2009).

1.2.3.3. Cathepsins and other serine proteases

Other important proteases in ECM remodeling belonging to the cathepsin family (CTSs). Even though most cathepsins are classified as cysteine proteases (CTSB/C/F/H/K/L/O/S/V/X/W), others may function as serine (CTSA and CTSG), or aspartic (CTSD and CTSE) proteases, (Fonovic & Turk 2014). CTSs size range around 20-30 kDa and, like most proteases, contain pre- and propeptide regions, and a catalytic domain with define substrate affinity based in the presence of a cysteine, serine or aspartate residue in the active site (Fonovic & Turk 2014; Bromme & Wilson 2011; Reiser et al. 2010). There are two main mechanisms for CTSs release in the extracellular space: altered traffic of newly formed enzyme (with overstimulation of trans-Golgi secretory pathway), or release from endosomes, lysosomes and azurophilic granules. CTSs may be synthetized by macrophages (CTSB, L, S, and K), mast cells (CTSL), fibroblasts, and also polymorphonuclear neutrophils (CTSG). Several CTSs are known to play key roles in the ECM remodeling of bronchial tissues, such as CTSK, L and S that have strong affinities towards elastin fibers, but are also capable

of collagen type IV and fibronectin proteolysis. Conversely, CTSK cleaves fibrillary collagens - types I and II (Fonovic & Turk 2014; Kasabova et al. 2011; Wolters & Chapman 2000). CTSC found in alveolar tissues is reported to activate other enzymes such as elastase and CTSG. (Kasabova et al. 2011).

Neutrophil elastase (ELANE) and protease 3 (PTRN3) are two other serine proteases with 29 and 33 kDa, respectively, that are normally stored in the azurophilic granules of polymorphonuclear neutrophils. These proteases are secreted upon neutrophil activation at inflammatory sites, by cleavage of the N-terminal peptide and removal of an aminoterminal dipeptide by CTSC (Korkmaz et al. 2010). Both serine proteases are capable of degrading various ECM structural molecules, including elastin, type IV collagen, and fibronectin. If dysregulated these molecules may compromise integrity of bronchial and alveolar walls. Importantly, ELANE and PTRN3 may also cleave inflammatory mediators, cell receptors and lung surfactant molecules with potential impact in ECM remodeling (Sinden & Stockley 2013; Lucas et al. 2011; Korkmaz et al. 2010).

1.2.3.4. Protease Inhibitors

Tissue metalloproteases inhibitors (TIMPs) are a small family of homologous proteins, mainly synthesized by connective tissue cells and leukocytes that control the activity MMPs, ADAMs and ADAMTSs by means of forming noncovalent complexes with their targeted proteases. Briefly, these proteins comprise two domains (N- and C-terminals) stabilized by three disulfide bonds, in which the N-terminal is the active domain containing two zinc ions, with one folding within itself to bind and inhibit metalloproteases (Mocchegiani et al. 2011; Rocco et al. 2001). Although all four TIMP may work as metalloprotease inhibitors, these have different regulatory efficiencies according to their best affinity to each protease. For example, whereas TIMP2 has a greater affinity to MMP-2, TIMP3 is a stronger inhibitor of MMP-9. Moreover, TIMP1 is capable of controlling the activity of most MMPs, except for some membrane type members (MMP-14-16, -19, -24). In this family, TIMP3 displays the most wide inhibitory range, being able to efficiently regulate a vast number of ADAMs (ADAM10, 12, 17, 28, 33) and ADAMTSs (ADAMTS1, 2, 4, 5) (Arpino et al. 2015; Brew & Nagase 2011). TIMP3 also differs from the other family members in its tissue placement, while it is attached to the ECM, the remaining TIMPs are present as soluble inhibitors (Reunanen & Kähäri 2013).

Notably, TIMPs, besides directly controlling ECM proteolysis through the regulation of metalloproteases, can also influence the ECM turnover by balancing concentration levels of both TIMPs and metalloproteases. Moreover, these proteases can release and activate sequestered TGFβ in the ECM, which in turn may lead to fibrosis, as TGFβ promotes matrix deposition. TIMPs, by

conditioning metalloproteases activity, impairs TGFβ release, regulating that deposition. Furthermore, TIMPs can also control ECM turnover, through the regulation of inflammatory pathways, avoiding the cleavage of cell-surface cytokines and cytokine receptors by metalloproteases (Arpino et al. 2015).

The serine proteases inhibitors (SERPINs) superfamily comprises at least 36 functional members, subdivided into distinct clades (A to I) according to similarities in protein sequence, gene organization and chromosomal location. All SERPINs share a highly conserved three-dimensional structure, characterized by a prototypical molecular arrangement in three β-sheets, nine α-helixes and an exposed reactive center loop (RCL; Fig. 1.1). This domain contains a pseudo-substrate (P1-P1') that once cleaved and covalently bound to target proteases, inhibits their activity in an irreversible fashion. Although SERPINs regulate mostly serine proteases, some are able to control the activity of cysteine proteases and importantly, SERPINs display different affinities toward different proteases, neutralizing not only specific enzymes but also wide classes of proteases (Seixas 2015; Gooptu & Lomas 2009). Alpha-1-antitrypin (SERPINA1), the major protease inhibitor in the serum, represents a critical regulator of ECM degradation in the lower respiratory tract by controlling the enzymatic activity of ELANE, but also CTSG and PRTN3. However, SERPINA1 is also a potential target for MMPs, which are capable of cleaving the RCL, rendering this molecule inactive (Fortelny et al. 2014).

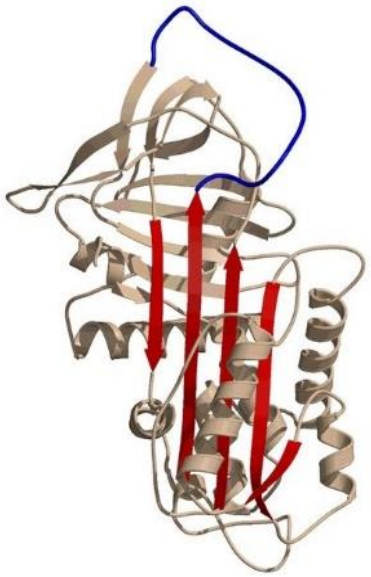

Figure 1.1. Conserved SERPIN three-dimensional structure. Archetypical SERPIN structure with reactive center loop highlighted in blue, and functional domain "shutter" in red [Adapted from (Gettins & Olson 2016)].

Clade B SERPINs are also particularly relevant in the control of lung ECM degradation by preventing cell death (apoptosis and necrosis) and by avoiding promiscuous proteolysis associated to the release of diverse proteases found in the lysosome and cytolytic granules (Table 1.3) (Sun et al. 2016; Houghton 2015; Moroy et al. 2012; Askew & Silverman 2008). SERPINE1 is another example of a protease inhibitor, which functions in lung ECM (alveolar space) by preventing fibrin deposition, an important event in fibrosis and acute lung injury (Askew & Silverman 2008).

Table 1.3. Clade B SERRPINs with known roles in lung function [Adapted from (Askew & Silverman 2008)].

SERPIN	Targets	Function
SERPINB1	Neutrophil elastase (ELANE) Cathepsin G (CTSG) Proteinase-3 (PRTN3)	Protection from elastase activity
SERPINB2	Urokinase-type plasminogen activator (PLAU) Tissue-type plasminogen activator (PLAT)	Protection against cell death
SERPINB3	Cathepsins K, L, S, V (CTSK, L, S, V)	Protection from cytosolic lysosomal peptidases; Inhibition of cell death
SERPINB4	Cathepsin G (CTSG) Mast cell proteinase (MCP)	Protection against cell death
SERPINB6	Cathepsin G (CTSG)	Protection from granule peptidases
SERPINB9	Granzyme B (GZMB)	Protect cytosolic lymphocytes; Protection against cell death
SERPINB10	Trypsin (PRSS1) Thrombin (F2)	Protection against cell death
SERPINB12	Trypsin (PRSS1)	
SERPINB13	Cathepsins K and L (CTSK, L)	Protection against cell death

1.3. Complex Lung Diseases

Complex diseases affecting the respiratory system can exhibit alterations of the ECM remodeling and concomitantly display dysregulated proteolytic processes that on one hand, can lead to lung parenchyma loss and cell death, and on the other, be correlated to cell proliferation and abnormal accumulation of fibrotic material and increased tissue stiffness. COPD and LC are two

examples of complex diseases that are linked to dysregulated ECM remodeling. Some of the phenotypes associated to these diseases are the pulmonary emphysema (in COPD) and the pulmonary fibrosis (in LC), both correlated with a dysregulation of ECM remodeling (Bidan et al. 2015; Cox & Erler 2011).

1.3.1. Chronic Obstructive Pulmonary Disease

Chronic Obstructive Pulmonary Disease (COPD) is the most common smoking-related disease in Western countries, which is predicted to represent the third cause of death by the end of 2020 (Vestbo et al. 2013). This illness is defined by the presence of progressive airflow obstruction that is not fully reversible and its major clinical manifestations are emphysema and chronic bronchitis (Fig.1.2) (Pauwels et al. 2012). Whereas the former COPD phenotype is characterized by an enlargement and destruction of alveoli, resulting in lower pulmonary oxygenation. The latter one is associated with inflamed and thickened bronchial walls, together with a luminal obstruction by mucus and inflammatory cells, normally causing breathing difficulties and persistent cough (Fischer et al. 2011).

Cigarette smoking is the primary environmental risk factor for COPD, but there are other elements such as exposure to noxious gas/vapours, biological and mineral dust that have also been linked to an increased COPD hazard (Matheson et al. 2005; Trupin et al. 2003). In general, all these environmental factors are thought to trigger oxidative stress, enhancing the levels of reactive oxygen species (ROS) in the respiratory tract, and consequently causing an inflammatory response with potential effects in the lung ECM (Rahman & Adcock 2006).

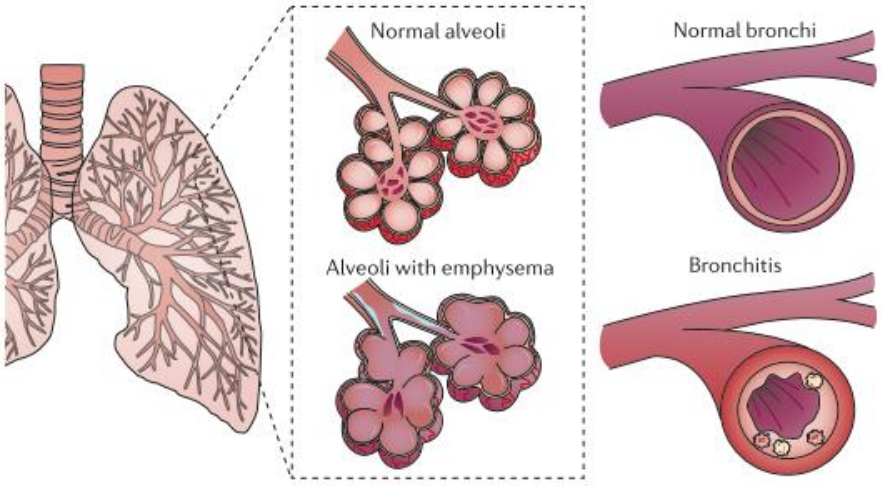

Figure 1.2. COPD Phenotypes. Emphysema and Chronic Bronchitis are the most common COPD phenotypes, characterized by thickening and consequent destruction of alveoli, and obstruction of airflow by mucus and inflammatory cells, respectively. [Image from (Houghton 2013)].

Spirometry is the "gold standard" for measuring the extent of airflow limitation. Forced expiratory volume in one second (FEV1), and forced vital capacity (FVC) are two lung function parameters assessed by spirometry that are fundamental to the evaluation of the degree of airflow limitation and its progress overtime. According to the guidelines of the Global Initiative on Chronic Obstructive Pulmonary Disease (GOLD), a health entity working with World Health Organization (WHO) in this field, a COPD diagnosis is confirmed if post-bronchodilator FEV1/FVC ratio is below 70%. In addition, GOLD also recognizes four COPD stages based in their severity, starting from Mild (GOLD 1; FEV1 ≥80%), passing through Moderate (GOLD 2; FEV1 50-79%) and Severe (GOLD 3; FEV1 30-49%), until reaching a Very Severe stage (GOLD 4; FEV1 <30%) (GOLD Guidelines 2017). Although a FEV1/FVC cutoff of <70% is widely used in COPD diagnosis it may be less accurate in elderly, and in adults under 45 years old, resulting in the first case in a overestimation, and in the second in a reduction in the numbers of affected subjects, particularly for milder phenotypes. This phenomenon has been attributed to regular alterations in lung function associated to ageing (Ito & Barnes 2009; Roberts et al. 2006).

1.3.2. Lung Cancer

Lung cancer is the most lethal cancer worldwide, and its incidence has increased significantly during the XX century, mainly due to a global raise in smoking exposure. Interestingly, recent studies are pointing to a reduction in the number of affected males, whereas female rates are being maintained constant (Ridge et al. 2013). Despite cigarette smoking has been considered the main environmental risk factor for LC, as in COPD, others inhaled smokes and particles (organic and inorganic) are recognized to play a role in the disease pathogenesis.

Two major types of LC can be identified: the small-cell lung cancer (SCLC) type, a rarer form of the disease observed in 15-20% of patients, and the non-small-cell lung cancer (NSCLC) reported in 80-85% of cases. NSCLC is itself further divided into three major histologic subtypes: squamous-cell carcinoma (SCC) and adenocarcinoma (ADC); the two most frequent subtypes of LC, and large-cell lung cancer (Bracci et al. 2012; Herbst et al. 2008). Notably, ADC usually has a slower growth than other LC types and it is more prevalent in females than males. Moreover, whereas ADC is more frequently detected in peripheral lung parenchyma, SCC, is more usually related to bronchial epithelial lesions and more commonly found in the central region of the lung. Also, both types are highly associated with smoking, although in non-smoking patients, ADC is the most common LC type (The American Cancer Society 2016; Chang et al. 2015).

LC is classified through different stages, based of tumor size (T1-4), lymph node involvement (N1-3) and presence of metastasis (M0-1) (Currie et al. 2009). Typically, LC is diagnosed when there is already an extensive cell proliferation and metastization to other areas of the body. In fact, about half of the patients display distant metastasis at the time of the diagnosis, resulting in a late stage LC diagnosis. The gold standard procedure for LC screening is a chest X-ray. In addition, suspected and confirmed LC patients are often submitted to bronchoscopy, needle biopsy of the lung, surgical procedures and/or an ultrasound (Field et al. 2013; Currie et al. 2009). However, new techniques are being evaluated for a better and early LC diagnosis, such as low-dose computed tomography (LDCT) and positron emission tomography with computed tomography (PET-CT), especially in groups at higher risk of LC (subjects over 55 yrs. and/or smokers) (Field et al. 2013; Humphrey et al. 2013).

1.4. Mechanistic links between COPD and LC

Over the latest years, it has been hypothesized that COPD and LC share a common mechanism in their pathogenesis (Houghton 2013; Vermaelen & Brusselle 2013; Young et al. 2011). First, COPD patients present a higher risk for LC (2-5 time higher). Second, not only, LC is a common complication in COPD, as COPD itself is also a prevalent co-morbidity in LC cases (Durham & Adcock 2015; Young et al. 2015). More precisely, in COPD, the emphysema phenotype has been pointed out as the stronger marker for LC risk. Typically, NSCLC (ADC and SCC), the most frequent cancer type among COPD, reaches about 80 to 85% of the cases, with a higher incidence of ADC, in comparison to SCC (Gabrielson 2006; Papi et al. 2004). Third, COPD and LC were found to overlap in several genetic susceptibility factors (see section 1.4.1. below). However, COPD has been reported to display stronger familial aggregation than LC, while COPD estimated heritability ranges from 40-75%, in LC it only reaches 15-25% (Young et al. 2012). Fourth, COPD and LC have as a major etiological factor cigarette smoking, as well as other inhaled elements, which may be correlated with increased fields of injury, chronic inflammation, enhanced oxidative stress (see section 1.4.2. bellow), and consequently with altered ECM remodeling (see section 1.4.3. below) (Houghton 2013).

1.4.1. Genetic Susceptibility factors

The recent efforts made by independent genome wide association studies (GWAS) to underpin common variants increasing COPD and LC susceptibility, uncovered in several instances the same candidate risk genes. For example, these included acetylcholinergic nicotinic receptors, subunits *α3* and *α5* (*CHRNA3* and *CHRNA5*, respectively) (15q25), hedgehog interacting protein (*HHIP*) (4q31), family with sequence similarity 13 member A (*FAM13A*) (4q24), and iron responsive element binding protein 2 (*IREB2*) (15q25) (Khiroya & Turner 2015; Yang et al. 2013; Young et al. 2011).

Notably, *CHRNA3*/*A5* association to COPD and LC has been replicated several times in independent cohorts and in distinct human groups, in European (Hardin et al. 2012; Marchand et al. 2009) and in Asians (Kim & Lee 2015). Besides a correlation with nicotinic addiction and smoking behavior mediated through an effect in neuronal cells, CHRNA3/A5 receptors have been proposed to have a more direct impact in disease progression and bronchial epithelial cells by inducing inflammation, with effects in cell proliferation rate, inhibition of apoptosis and malignant

transformation (Dang et al. 2016; Singh et al. 2011). Conversely, HHIP, which regulates the activity of Sonic Hedgehog (SHH), has been reported to have a critical role in the signaling pathways for both bronchial embryogenesis and lung development. Also, in COPD and LC, it has been proposed to participate in the cycles of injury and repair induced by environmental risk factors, in which effects in epithelial mesenchymal transition (EMT) can further contribute to LC pathogenesis, by allowing cells to increase their motility (Kugler et al. 2015). Regarding *IREB2*, a gene located in the same chromosomic region of *CHRNA3/A5* (15q25), its association to COPD and LC has been hypothesized to be connect to iron regulatory pathways and *IREB2* role in iron homeostasis (Ziółkowska-Suchanek et al. 2015; Alder et al. 2011). The function of *FAM13A*, not completely understood, it is thought to be correlated to signal transduction, due to described effects in tumor cell migration and possible impact in cancer growth (Eisenhut et al. 2017; Ziółkowska-Suchanek et al. 2015; Cho et al. 2010).

Still, to date, SERPINA1 deficiency (AATD) remains one of the few proven genetic causes for emphysema, mainly due to a pathogenic variant (rs28929474; p.Glu342Lys) leading to the unopposed activity of ELANE, cleavage of elastin fibers and ECM degradation. However, AATD only accounts for a small proportion of COPD cases (1-3%) and in LC it is not yet clear if it has indeed a role in tumorigenesis. In most GWAS for COPD and LC, no strong association of *SERPINA1* to disease susceptibility was detected, although it was described a moderate association of rs28929474 variant with severe airflow limitation in COPD (Jackson et al. 2016; Enewold et al. 2012; Denden et al. 2010).

Sequence variation of MMPs has been extensively investigated in the context of COPD and LC by independent candidate gene approaches. Most interesting associations to lung disease include *MMP12* variants rs652438 and rs2276109 that were associated to emphysema, and severe stages of COPD (GOLD III-IV) and in *MMP2*, rs243865 variant, which has been linked to a decay in survival time of NCSLC patients (González-Arriaga et al. 2012; Haq et al. 2010).

1.4.2. Oxidative stress, cell injury and inflammation

A continuous exposure of the respiratory system to cigarette smoking and/or other occupational hazards is known to cause an increment in reactive oxygen species (ROS) and other chemical particles, leading to significant molecular changes, such as protein oxidation and DNA methylation (Yoshida & Tuder 2007), while also triggering cell injury and several pro-inflammatory responses (Bowler et al. 2004). Indeed, COPD and LC patients frequently show lungs with infiltrates of inflammatory cells, such as macrophages, neutrophils and monocytes, which can also release ROS

into lung microenvironment. In COPD, these radicals, may additionally cause the inactivation of important proteases inhibitors, resulting in an exacerbation of neutrophil elastase activity, loss of lung elasticity, apoptosis and emphysema (John et al. 2017; Domej & Oettl 2014). On the other hand, in LC individuals, ROS are often implicated in protein degradation and DNA methylation, later contributing to cancer development through the activation of anti-apoptotic molecules that increase cell division and proliferation, and facilitate tumor metastasis (Liou & Storz 2010).

Another outcome of the chronic inflammation in both COPD and LC microenvironments is the perpetuation of tissue injury (repeated cycles of tissue injury and repair) (Vakkila & Lotze 2004). Briefly, lung injuries are initiated by exogenous factors, like cigarette smoking. Then, injured cells start releasing diverse repair-linked mediators such as different families of epidermal and fibroblast growth factors (e.g. TGFα, HGF), chemokines, interleukins and prostaglandins. Later, these molecules are expected to impact ECM remodeling, by processes such as mitosis, migration and repair stimulation, involving collagen, laminin, fibronectin and matrix-metalloproteases, such as MMP-1 and -9 (Crosby & Waters 2010). Conversely, several repair-linked mediators also trigger the recruitment of macrophages and neutrophils to sites of injury that secret diverse proteases capable of degrading ECM elastin and collagens. Importantly, fragments derived from ECM proteolysis can also act as repair-linked mediators, directly or indirectly, supporting further chronic inflammation (Bonnans et al. 2014; Shifren & Mecham 2006). Moreover, in both diseases, the persistence of cell injury is believed to change cell death from apoptosis to necrosis. Contrary to apoptosis, a programmed cell death in which the cells usually shrink and maintain integrity of their membrane, in necrosis, cells lose that integrity and leak their internal contents to extracellular space, promoting an inflammatory response (Rock & Kono 2008).

Furthermore, important molecular changes can occur with chronic inflammation including an increase in DNA replication, angiogenesis, fibrotic processes and suppression of adaptive immunity that altogether facilitate the arisen of cancer cells with several genomic abnormalities, tumor growth and survival, and later metastization (Fig 1.3) (Vermaelen & Brusselle 2013).

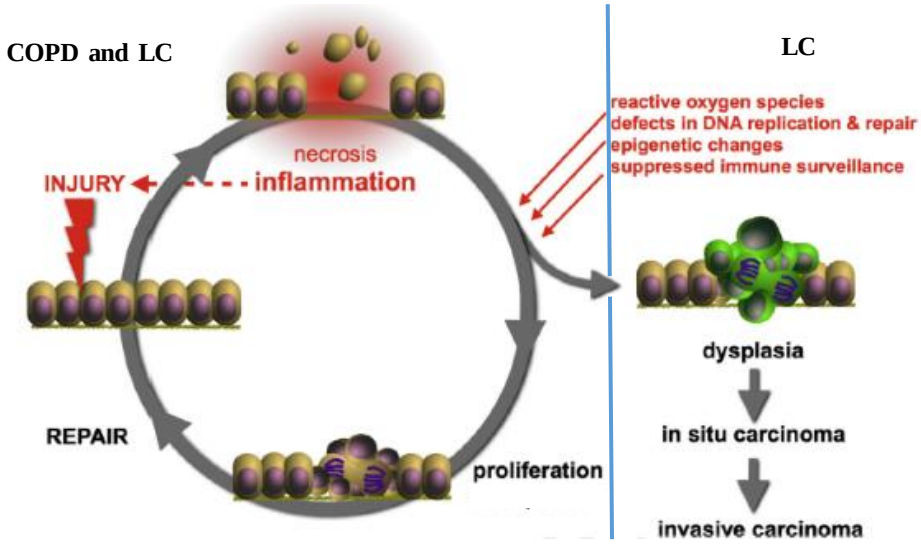

Figure 1.3. Repetitive cycles of tissue injury and repair. These cycles contribute to a state of chronic inflammation, a feature found in both COPD and LC, which consequently may lead to malignant degeneration. Eventually, inflammatory mediators resulting from this cycles can induce genetic aberrations, perpetuated by the accumulation of cells that escape the apoptotic process, resulting in dysplasia, followed by carcinoma in-situ, and finally an invasive carcinoma state, developing LC [Adapted from (Vermaelen & Brusselle 2013)].

1.4.3. ECM remodeling and proteolysis in lung disorders

In lungs injured by chronic inflammation is common to observe a disequilibrium between collagen expression, deposition and turnover, with an overall augmented content of collagens in comparison to elastin. In mild and moderate COPD cases, collagen deposition is thought to contribute to the thickening of bronchial walls associated to the development of chronic bronchitis (Eurlings et al. 2014; Annoni et al. 2012; Harju et al. 2010; Kranenburg et al. 2006). In a LC situation, the overexpression and accumulation of collagens type I and III is regarded as an important factor for tumor stroma stiffening and cancer microenvironment (Burgstaller et al. 2017; Vicary et al. 2017; Burgess et al. 2016).

Proteolytic imbalance has been proposed as one of the major causes for COPD as it is occurs in pulmonary emphysema associated to AATD. Still, other studies have implicated proteolytic imbalance in the pathogenesis of COPD and LC in connection mainly to MMPs activities. For example, MMP-1 that was associated with both diseases through different genetic studies, was suggested to promote the metastasis formation in LC, through interaction with STAT3 (Schütz et al. 2015) and described to cause airways enlargement and emphysema, when overexpressed in lungs, with cigarette smoke and other noxious gases exposure, and inflammation as key elements (Churg et

al. 2012; Mocchegiani et al. 2011; Greenlee et al. 2007). Furthermore, MMP-12, which is considered an important factor for COPD severity, was found to be highly expressed in patients alveolar macrophages and reported to induce emphysema driven by cigarette smoking (Churg et al. 2012; Soto-quiros et al. 2009). Noticeably, increased levels of MMP-12 were also associated with TIMP1 hydrolysis, SERPINA1 inactivation and increased ELANE activity (Houghton 2015; Lucas et al. 2011).

In LC conditions, a large diversity of MMPs are known to facilitate tumor growth and invasiveness through their impact in the degradation of ECM barriers and as promotors of angiogenesis. Furthermore, MMPs are also known to favor metastization, given that MMPs are critical molecules in ECM cell detachment, thus allowing cancer cells to enter in circulation and to reach distant tissues (Reunanen & Kähäri 2013). In overview, significant changes in the ECM composition associated to its proteolytic degradation and/or wound-healing (fibrosis) can be correlated with the outcomes of COPD and LC.

1.5. Patient Tailored Therapeutics

To date, patient tailored therapeutics are already available for NSCLC. In particular, for subjects carrying specific somatic mutations in genes such as the Epidermal Growth Factor Receptor (*EGFR*), Kirsten Rat Sarcoma Viral Oncogen Homolog (*KRAS*), or Anaplastic Lymphoma Kinase (*ALK*) genes (Ridge et al. 2013). Even thought, some of these therapies for ADC and SCC were already proven to be very successful in some cases, these cannot be administrated widely as most patients lack corresponding molecular targets not showing any response to treatment. Therefore, most patients continue to be medicated with standard drugs with lower response rates and frequent side-effects (Cortinovis et al. 2016; Lazarus & Ost 2013).

In COPD, in spite of a large spectrum of therapeutics available very few take into account patient molecular profile. The only exception is SERPINA1 augmentation therapy in pulmonary emphysema associated to AATD. In near future, possibly some drugs targeting β2-adrenegic receptor (*ADRB2*) mutations, may promote muscle relaxation and dilation of airways (Nielsen et al. 2017; Wewers & Crystal 2013).

MMPs and several serine proteases have been investigated as possible therapeutic targets in lung diseases, but so far, there are no perspective of these studies being translated soon into clinical practice (Moroy et al. 2012). Interestingly, somatic mutations in two inhibitor genes, *SERPINB3* and *SERPINB4*, displaying high sequence similarities but divergent protease affinities (see Table 1.3),

were associated with a better outcome in melanoma treatment with anti-CTLA4 antibody (ipilimumab) (Riaz et al. 2016). This drug, ipilimumab, is currently undergoing phase III trials for LC treatment (Bustamante Alvarez et al. 2015), positioning *SERPINB3* and *SERPINB4* as promising targets for personalized medicine in LC.

AIMS

In this study, we explore the impact of proteolysis related genes, including proteases, their inhibitors and lung ECM substrates in two correlated lung disorders: COPD and LC that share common and distinctive features of ECM degradation and remodeling. To achieve this main goal, we used different methodological approaches to address the following specific objectives:

1. Survey of The Cancer Genome Atlas (TCGA) database for available clinical and epidemiological data of two LC subtypes (ADC and SCC);
2. Evaluation of TCGA database regarding the mutational (somatic and germline) landscape of 73 proteolysis candidate genes with function in lung ECM;
3. Analysis of the sequence variability of selected genes (*SERPINB3/B4* and *CTSG*) in a small cohort of COPD and LC patients of Portuguese origin.

2. Materials and methods

2.1. Bioinformatics analysis

2.1.1. TCGA data – Lung Cancer

Clinical and epidemiological information, as well as sequence variation (somatic and germline mutations) and expression data for 522 ADC and 504 SCC patients were retrieved from The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/). For LC samples, TCGA provides several important clinical and epidemiological variables such as tumor histology, anatomic site (upper, middle, lower, right and left lung), age at initial diagnosis, pulmonary function tests (FEV1 and FVC), smoking history (packs per year - PPY), population of origin (United States, Europe, Vietnam, Australia, and Canada), ethnicity (white, black, Hispanic and non-Hispanic) and gender. However, a considerable number of samples was found to lack information for some variables like smoking history and pulmonary function tests (FEV1 and FVC), which in the latter case impaired the identification of COPD burden among ADC and SCC.

TCGA sequence variation data results from whole-exome or whole-genome sequencing (WES and WGS) of both tumor and non-tumor tissues (lung healthy section or blood). In this situation, reads assembling and variant calling were performed through a direct comparison of tumor and non-tumor results using MuTect2 software. Briefly, this software combines several Bayesian methods in the filtering of low-quality sequence data; variant detection in tumor samples; removal of false positives; and in the discrimination of somatic and germline variants (Cibulskis *et al*, 2013). In the analysis of matched tumor/normal samples, somatic mutations are only found in tumors, while germline mutations can be present in both samples or if found in tumor their status could not be evaluated. Expression data provided by TCGA, is derived from RNA sequencing.

2.1.2. Clinical and epidemiological data analysis

Several statistical analyses were carried out with selected variables of TCGA to address the impact of different risk factors in ADC and SCC onset. Specifically, we considered for our

comparisons within and between LC subtypes, the variables ethnicity and population of origin combined as European ancestry, African-Americans and Asians, age at diagnosis (mean values), gender, tumor anatomic site, and smoking history (mean PPY values). The Mann-Whitney implemented through MedCalc software (version 17.6) was used to appraise statistical significance of all set of comparisons done.

2.1.3. Candidate genes selection

For this study, we selected 73 proteolysis related candidate genes (table 2.1) based in three main criteria: 1) evidence for the occurrence in lung tissues based either in gene expression data or proteomic screenings of biological samples withdrawn from lungs, such as bronchoalveolar lavage fluid (BALF) or sputum (SP) (Ohlmeier et al. 2012; Casado et al. 2007; Plymoth et al. 2006); 2) known function in ECM remodeling (ECM organization – degradation of the ECM) as annotated in the Reactome pathway database (http://reactome.org/); and genes found to be associated to either COPD and/or LC according to the GWAS catalogue (http://www.ebi.ac.uk/gwas/).

Table 2.1. Proteolysis related candidate genes selected for this study.

Gene ID	Name	Proteolysis activity
A2M	ALPHA-2-MACROGLOBULIN	serine protease inhibitor
ADAM15	DISINTEGRIN AND METALLOPROTEINASE DOMAIN-CONTAINING PROTEIN 15	metalloprotease
ADAM17	DISINTEGRIN AND METALLOPROTEINASE DOMAIN-CONTAINING PROTEIN 17	metalloprotease
ADAM19	DISINTEGRIN AND METALLOPROTEINASE DOMAIN-CONTAINING PROTEIN 19	metalloprotease
ADAM9	DISINTEGRIN AND METALLOPROTEINASE DOMAIN-CONTAINING PROTEIN 9	metalloprotease
ADAMTS1	DISINTEGRIN AND METALLOPROTEINASE WITH THROMBOSPONDIN MOTIFS 1	metalloprotease
ADAMTS8	DISINTEGRIN AND METALLOPROTEINASE WITH THROMBOSPONDIN MOTIFS 8	metalloprotease
CAPN1	CALPAIN-1 CATALYTIC SUBUNIT	cysteine protease
CAPN2	CALPAIN-2 CATALYTIC SUBUNIT	cysteine protease
CAPNS1	CALPAIN SMALL SUBUNIT 1	cysteine protease
CAST	CALPASTATIN	cysteine protease inhibitor
COL14A1	COLLAGEN XIV ALPHA 1 CHAIN	Collagen (substract)
COL1A1	COLLAGEN TYPE I ALPHA 1 CHAIN	Collagen (substract)
COL1A2	COLLAGEN TYPE I ALPHA 2 CHAIN	Collagen (substract)

COL3A1	COLLAGEN TYPE III ALPHA 1 CHAIN	Collagen (substract)
COL4A1	COLLAGEN TYPE IV ALPHA 1 CHAIN	Collagen (substract)
COL4A2	COLLAGEN TYPE IV ALPHA 2 CHAIN	Collagen (substract)
COL4A3	COLLAGEN TYPE IV ALPHA 3 CHAIN	Collagen (substract)
COL5A2	COLLAGEN TYPE V ALPHA 2 CHAIN	Collagen (substract)
COL6A1	COLLAGEN TYPE VI ALPHA 1 CHAIN	Collagen (substract)
COL6A2	COLLAGEN TYPE VI ALPHA 2 CHAIN	Collagen (substract)
COL6A3	COLLAGEN TYPE VI ALPHA 3 CHAIN	Collagen (substract)
COL8A1	COLLAGEN TYPE VIII ALPHA 1 CHAIN	Collagen (substract)
CST3	CYSTATIN-C	cysteine protease inhibitor
CST6	CYSTATIN-M	cysteine protease inhibitor
CTSB	CATHEPSIN B	cysteine protease
CTSD	CATHEPSIN D	aspartic protease
CTSG	CATHEPSIN G	serine protease
CTSK	CATHEPSIN K	cysteine protease
CTSL1	CATHEPSIN L1	cysteine protease
CTSS	CATHEPSIN S	cysteine protease
ELANE	NEUTROPHIL ELASTASE	serine protease
ELN	ELASTIN	elastin (substract)
EMILIN2	EMILIN-2	elastin (substract)
FN1	FIBRONECTIN	fibronectin (substrate)
FURIN	FURIN	serine protease
KLK1	KALLIKREIN-1	serine protease
KLK4	KALLIKREIN-4	serine protease
KLKB1	PLASMA KALLIKREIN	serine protease
LAMA3	LAMININ SUBUNIT ALPHA 3	Laminin (substrate)
LAMA5	LAMININ SUBUNIT ALPHA 5	Laminin (substrate)
LAMB1	LAMININ SUBUNIT BETA 1	Laminin (substrate)
LAMB3	LAMININ SUBUNIT BETA 3	Laminin (substrate)
LAMC1	LAMININ SUBUNIT GAMMA 1	Laminin (substrate)
LAMC2	LAMININ SUBUNIT GAMMA 2	Laminin (substrate)

MMP1	INTERSTITIAL COLLAGENASE	metalloprotease
MMP10	STROMELYSIN-2	metalloprotease
MMP12	MATRIX METALLOPROTEINASE-12	metalloprotease
MMP13	COLLAGENASE 3	metalloprotease
MMP14	MATRIX METALLOPROTEINASE-14	metalloprotease
MMP15	MATRIX METALLOPROTEINASE-15	metalloprotease
MMP2	72 KDA TYPE IV COLLAGENASE	metalloprotease
MMP3	STROMELYSIN-1	metalloprotease
MMP8	NEUTROPHIL COLLAGENASE	metalloprotease
MMP9	MATRIX METALLOPROTEINASE-9	metalloprotease
PI3	ELAFIN	serine protease inhibitor
PLG	PLASMINOGEN	serine protease
PRTN3	MYELOBLASTIN	serine protease
SERPINA1	ALPHA-1-ANTITRYPSIN	serine protease inhibitor
SERPINA3	ALPHA-1-ANTICHYMOTRYPSIN	serine protease inhibitor
SERPINB1	LEUKOCYTE ELASTASE INHIBITOR	serine protease inhibitor
SERPINB3	SQUAMOUS CELL CARCINOMA ANTIGEN-1	serine protease inhibitor
SERPINB4	SQUAMOUS CELL CARCINOMA ANTIGEN-2	serine protease inhibitor
SERPINB6	PLASMINOGEN ACTIVATOR INHIBITOR 2	serine protease inhibitor
SERPINC1	ANTITHROMBIN-III	serine protease inhibitor
SERPINE1	PLASMINOGEN ACTIVATOR INHIBITOR 1	serine protease inhibitor
SERPINE2	GLIA-DERIVED NEXIN	serine protease inhibitor
SERPING1	PLASMA PROTEASE C1 INHIBITOR	serine protease inhibitor
SLPI	ANTILEUKOPROTEINASE	serine protease inhibitor
THSD4	THROMBOSPONDIN TYPE-1 DOAMIN-CONTAINING PROTEIN 4	metalloendopeptidase
TIMP1	METALLOPROTEINASE INHIBITOR 1	metalloprotease inhibitor
TIMP2	METALLOPROTEINASE INHIBITOR 2	metalloprotease inhibitor
TIMP3	METALLOPROTEINASE INHIBITOR 3	metalloprotease inhibitor

2.1.4. Variants expression and impact analysis

In a first step, we examined the data available at the cBioPortal (http://www.cbioportal.org/), a comprehensive web tool that enables the analysis of large-scale cancer genomics datasets, including TCGA. This tool was used to get a glimpse of somatic mutation rates per each candidate gene.

In a more detailed analysis, VCF files containing all SNVs and INDELs (somatic and germline) identified by TCGA consortium were downloaded from the database. Then, these files were filtered for candidate genes genomic regions using *Tabix* software (version 0.2.6). Variants were compiled with *VCFTools* software version 4.0 (http://vcftools.sourceforge.net/), to remove those mutations with lesser quality (<20 reads).

Filtered somatic and germline variants for the 73 proteolysis candidate genes were next submitted to wANNOVAR software (http://wannovar.wglab.org/) analysis. This web tool has the advantage of compiling the results for a wide number of algorithms, predicting sequence variants functional consequences (SIFT, PolyPhen, CADD) together with variant frequencies for different human populations sequenced by large consortium like 1000 Genomes and ExAc. Here, we choose to consider only variant prediction effects of PolyPhen, SIFT and CADD algorithms. Whereas PolyPhen variant predictions are mainly based in protein sequence and structure, SIFT takes into account levels of evolutionary conservation, and CADD incorporates different metrics regarding functional data (not only protein structure) and conservation, prioritizing deleterious and pathogenic variants across wide range of functional categories (Eilbeck et al. 2017; Richards et al. 2015). Since CADD generates quantitative values, we used a cutoff of ≥14.5 in scaled CADD score, to denote most likely deleterious variants.

Candidate gene expression levels were obtained through Firebrowse (http://firebrowse.org/) engine, a TCGA online tool offering a direct comparison of expression differences between tumor and non-tumor samples.

2.2. Screening of Portuguese COPD and LC cases

Taking into account the results of our bioinformatics analysis we chose a few candidate genes for laboratory evaluation of sequencing variation in COPD and LC cases. Precisely, the selected genes for follow-up studies in our cases were *SERPINB3* and *SERPINB4* homologs and *CTSG*, which is regulated by *SERPINB4* activity.

2.2.1. Samples

Our sample collection included genomic DNA for COPD cases (N=43) sent to our laboratory for the AATD diagnosis and broncho-alveolar lavage fluid (BALF) of LC patients (N=45) gather in the scope of collaborative projects with clinicians from *Hospital São João* and *CEDOC researchers*. Our LC cases included 18 ADC and 2 SCC, for the remaining cases it was not possible to obtain NSCLC subtyping or were classified as belonging to other LC. However, not every sample was completely sequenced.

2.2.2. DNA Extraction

All samples derived from AATD diagnosis were previously extracted from blood using standard salting out methods or using *Generation Capture Column kit* (QIAGEN).

For BALF samples, DNA extraction was previously done using *QIAamp mini kit* (QIAGEN), according to manufacture instructions.

2.2.3. Polymerase Chain Reaction (PCR) amplification and sequencing

For COPD samples, which were found to have higher DNA concentrations, *SERPINB3* and *SERPINB4* genes were amplified in five different PCR reactions fragments using the primer pairs listed in Table 2.2. As high homologous genes, similar experimental schemes were used for *SERPINB3* and *SERPINB4* amplification, as schematized in Figure 2.1.

Briefly, fragment A spanning over 2.5 kb contained exons 1 to 3; fragments B and E ranged about 1.5 kb each and included exons 4-5 and exon 8, respectively, and finally, fragments C and D with 400 bp each, covered the remaining exon 6 and 7. All reactions were performed with the following reagents: 1*x KAPA Taq ReadyMix* or 1x *MyTaq Mix* and 0.5-1µM concentrations for each primer and 10-200 ng of genomic DNA, using the primers described in Table 2.2 and cycle conditions in *Annex Table T1*.

Table 2.2. Primers used for the amplification of *SERPINB3/B4* genes.

Gene(s)	Fragments	Primers
SERPINB3/B4	**A**	FW (B3): 5'- TGCTAAATGGAAGGACCACCAA -3' FW (B4): 5'- TGCTAAACAGAAGGACCATTGA -3' RV: 5'- CACTCTGTATGTCTCAATCT –3'
	B	FW: 5'- ACAGACTTAGCATGGGTTTA -3' RV (B3): 5'- TGTGATAATCCCTGCAGAACTTGT -3' RV (B4): 5'- TGTGATAATCCCTGCAGAACACAT –3'
	C	FW: 5'- TGGTCAGTGAGTCTAACAAT -3' RV (B3): 5'- TCATTAACTATGCCTTCAGTT -3' RV (B4): 5'- CAGAAATGTTTAACATTCCA -3'
	D	FW (B3): 5'- AATTTAAACATTTCTGATGGAATG -3' FW (B4): 5'- TAATATGTTAATACATGGAATGT -3' RV: 5'- AATATGAAGGTGAGTCATCA -3'
SERPINB3	**E**	FW: 5'- TGACACATGTAGTAGGCTGT -3' RV: 5'- CTTTCCCTTTCCAGAGAGAAAATG -3'
SERPINB4	**E**	FW: 5'- TGACACATGTAGTAGGCTGT -3' RV: 5'- TGCCCTTTCCAGAGAGAAAACAG -3'

For *CTSG* amplification a single PCR reaction was carried out to cover all five exons in a ~3 kb fragment. The sequences of the primers used are: Fw: 5'- TGAAACCTTTCATGGTAGCA -3' and Rv: 5'- GATCTTAGACTTCTTAGCCTCT -3'. PCR reaction mix consisted of 1*x KAPA Taq ReadyMix* or *MyTaq Mix*, 0.75 µM concentrations for each primer, and 15-150 ng of genomic DNA. The cycle conditions were the following: an initial denaturation step of 5 min at 95 ºC; 35 cycles of 30s at 94 ºC, 30s at 53 ºC, and 3 min at 68 ºC; and a final extension step of 20min at 68 ºC.

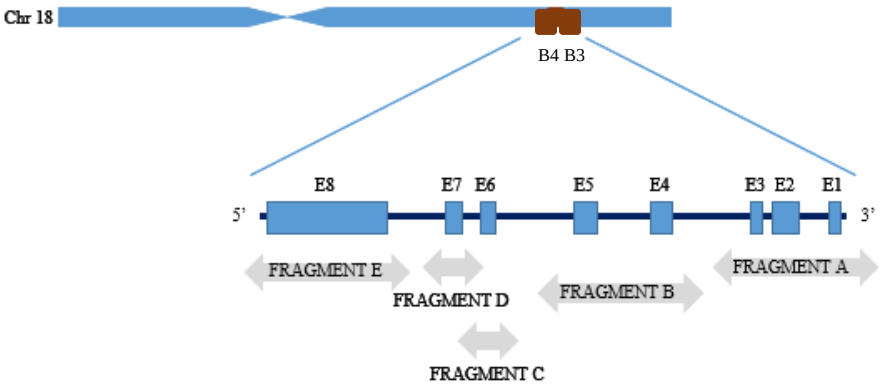

Figure 2.1. Schematic representation of *SERPINB3/SERPINB4* amplification. Location of *SERPINB3* (B3) and *SERPINB4* (B4) genes in chromosome 18 is shown on top (marked as red). The common gene structure is shown below with the amplicons represented as grey arrows.

In lung cancer samples, due to lower DNA concentrations from BALFs, *SERPINB3* and *SERPINB4* were amplified using semi-nested PCR reactions for fragments A, B and E. After a first round of PCR reactions using primer pairs listed in Table 2.2, LC samples were submitted to second round PCR using the primers listed in Table 2.3. Similarly to PCR reactions used in first round amplification, semi-nested mixtures were done using the following reagents: 1*x KAPA Taq ReadyMix* or *MyTaq Mix*, 0.5-1µM concentration for each primer, and 2 µL of the first PCR product diluted 1:50. PCR conditions of semi-nested reaction are shown in *Annex Table T2*.

Table 2.3. Semi-nested PCR primers used for *SERPINB3/B4* amplification.

Gene(s)	Fragments	Primers
SERPINB3/B4	**A**	FW: 5'- AGGAGAAGGCAATAGAATCC -3' RV: 5'- CACTCTGTATGTCTCAATCT –3'
	B	FW: 5'- ACAGACTTAGCATGGGTTTA -3' RV: 5'- CTGTGATTTCCTCCTTGGCT -3'
	E1	FW: 5'- TGACACATGTAGTAGGCTGT -3' RV: 5'- TGGGCTTATTAAGAGAAAGA -3'
	E2	FW: 5'- AGACCAACAGCATCCTCTTCT -3' RV (B3): 5'- CTTTCCCTTTCCAGAGAGAAAATG -3' RV (B4): 5'- TGCCCTTTCCAGAGAGAAAACAG -3'

For *CTSG* amplification in LC samples, there was no need for a second round of PCR reactions as the first ones provided satisfactory results.

All PCR reactions were evaluated by DNA electrophoresis in a 1.5% agarose gel, using SGTB (GRISP) commercial buffer and *GreenSafe Premium* (Nzytech) as a staining dye for DNA. Then, PCR products were visualized by Gel Doc XR+ System (Biorad).

For sequencing of target regions, amplified PCR products were first purified by column centrifugation (800 g) with *Sephacryl S-300 High Resolution (GE Healthcare)* resin. Next, sequencing reactions were carried using BigDye Terminator Sequencing version 3.1 cycle sequencing chemistry (Applied Biosystems) and specific primers (0.25 µM) listed in Table 2.4 covering *SERPINB3*, *SERPINB4* or *CTSG*. All sequencing reactions were carried using the cycling conditions described in *Annex Table T3*. Later, sequencing fragments were purified by column centrifugation (2000 g), with *Sephadex G-50 Fine DNA Grade* (GE Healthcare) resin. Finally, *Hi-Di Formamide* (Applied Biosystems) was added to all purified samples and electrophoretic analysis of sequencing products was carried out in an ABI3130 automated sequencer (Applied Biosystems).

Table 2.4. Primers used for SERPINB3/B4 and CTSGsequencing .

Gene(s)	Exons	Primers
SERPINB3/B4	**1**	FW: 5'- AGGAGAAGGCAATAGAATCC -3'
	2, 3	RV: 5'- CACTCTGTATGTCTCAATCT –3'
	4	RV: 5'- TCCCTAAATCCACACTTCAGT -3'
	5	FW: 5'- GAAGCAATGAATCTCCTTCA -3'
	6	FW: 5'- TGGTCAGTGAGTCTAACAAT -3'
	7	RV: 5'- AATATGAAGGTGAGTCATCA -3'
	8a	RV: 5'- TGGGCTTATTAAGAGAAAGA -3'
	8b	FW: 5'- AGACCAACAGCATCCTCTTCT -3'
CTSG	**1**	RV: 5'- CTGTATTCTTACCTCCTAGGTA -3'
	2, 3	FW: 5'- CATCTTCCAGCCTTTCTGGA -3'
	4	RV: 5'- ATGGAATCTGTTCGCACTGCCT -3'
	5	FW: 5'- TAAGGCAGAGCTGAAGTCCA -3'

2.2.4. Sequence analysis

All sequences were aligned to human reference sequence using *Geneious*, version 5.5.9, software. Precisely, the genomic segments containing *SERPINB3*, *SERPINB4*, and *CTSG* genes were retrieved from the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/). The coordinates of these segments, from GRCh37 assembly are 61322431 to 61329197 for *SERPINB3*, 61304493 to 61311532 for *SERPINB4*, and 25042728 to 25045466 for *CTSG*. All SNVs identified in our study were compared with data from dbSNP 150 available through the *University of California Santa Cruz (UCSC) Genome Browser* (https://genome.ucsc.edu).

2.2.5. Statistical Analysis

To evaluate if the SNVs identified in COPD and LC samples could differ in their frequencies from random populations we used the Iberian population (IBS) from 1000 Genomes (http://www.internationalgenome.org/data/) as control. For common variants, as defined by IBS MAF ≥ 0.05, Fisher's exact test was carried out to test possible associations of SNV to the disease. For low-frequency variants (MAF<5% in IBS), we used C-alpha test and burden test implemented to PLINK/SEQ package, version 0.10, (https://atgu.mgh.harvard.edu/plinkseq/) to detect any enrichment of deleterious variants (missense, UTR and splice region). Several sets of comparisons were performed to take into account possible associations to lung disease. We compared not only COPD and LC to the IBS population, but also the two diseases against each other and both diseases together against IBS.

3. Results and Discussion

3.1. TCGA data analysis

3.1.1.1. Epidemiological analysis

The "clinical data" file provided by TCGA allowed the evaluation of epidemiological information for two major non-small cell lung cancer (NSCLC) subtypes: adenocarcinoma (ADC) and squamous cell carcimoma (SCC). More specifically, in this study, we analyzed for a large dataset of ADC and SCC patients the following variables: population ethnicity (European, African, Asian and Native American ancestry), gender, smoking history, age at initial diagnosis, and anatomical site (see also 2.1.1 section).

In Table 3.1 the distribution of TCGA cases per each human ancestry groups, (European, African, Asian and Native-American) is shown. No differences were detected in the prevalence of ADC and SCC across major human groups (population ancestry) using the TCGA dataset. Even thought, this LC panel might be biased by a sample collection carried out mainly in developed and industrialized countries, such as the United States, Canada, Germany and Australia, where subjects of European descend are likely to represent the largest ancestry group, the percentage of LC cases in individuals of African origin is notable, considering that this group represent approximately 13% of the population of the United States (U.S. Census Bureau 2012). Consistently, African-Americans living in the United States have been described to display higher LC incidences than European-Americans and several factors have been hypothesized to explain such differentiation, from a greater exposure of African-Americans to tobacco smoking and a preference for menthol cigarettes, to an excessive exposition to occupational carcinogens and other sociocultural factors, as well as their underlying genetic susceptibility (Alberg et al. 2005; Higgins et al. 2003; Stewart IV 2001).

Table 3.1. Distribution of lung cancer subtypes ADC and SCC per patient population ancestry.

Population Ancestry	Adenocarcinoma (ADC) Frequency (N)	Squamous cell carcinoma (SCC) Frequency (N)
European	0.749 (391)	0.694 (350)
African	0.10 (52)	0.062 (31)
Asian	0.015 (8)	0.018 (9)
Native-American	0.002 (1)	0
Non-Available	0.134 (70)	0.226 (114)
Total number of cases	**522**	**504**

N- Absolute number of cases

According to the World Health Organization (WHO) the LC incidence is extremely reduced in African regions, such as Western, Middle and Eastern Africa, when compared to high rates of LC observed in North America, European countries and Eastern Asia (Fig. 3.1). Therefore, this distribution may be then interpreted as an evidence for a strong influence of environmental factors in more developed regions into LC incidence and mortality, such as smoking, air pollution, and occupational hazards, than the subject genetic ancestry as European or African.

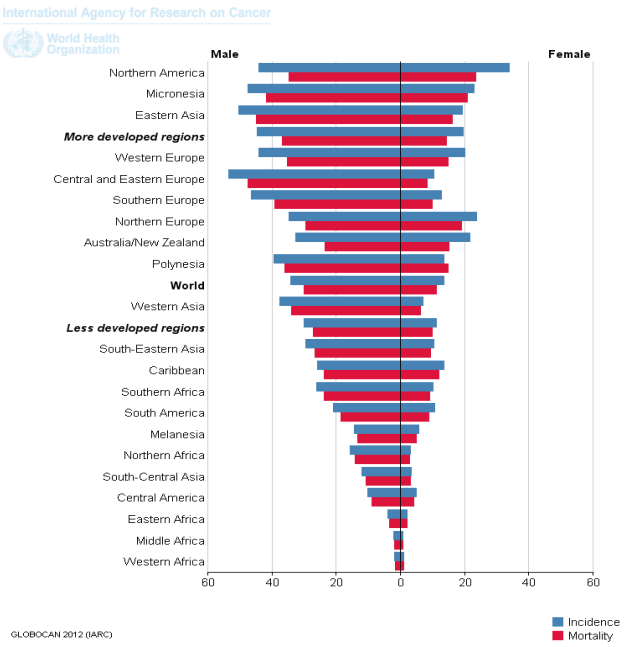

Figure 3.1. Lung cancer incidence and mortality rates by geographical populations and gender (2012 data). More developed societies tend to display higher LC incidence and mortality rates than less developed ones. Figure retrieved from Globocan webpage (http://globocan.iarc.fr/Default.aspx).

As illustrated in Figure 3.1 by WHO data, LC incidence and mortality is considerably higher in males than females. Again here, sociocultural factors associated to tobacco smoking are accepted to explain the actual gender differentiation in LC prevalence worldwide. However, in the near future and if considered a 30-year gap between the current smoking patterns and LC onset, the current observed gender difference is expected to be reduced and LC incidence to be identical between sexes, mainly due a decline of male smokers over the last decades (North & Christiani 2014; Alberg et al. 2005).

In this respect, the TCGA analysis did not disclose any statistical significant results for LC gender differences, neither in ADC or SCC (Table 3.2). Nevertheless, a trend for a higher incidence of SCC in males can be observed in Europeans and in the full dataset as a whole. There has been some controversy regarding the existence of specific risk factors from one gender or the other, but to date, most authors seem to agree that the higher incidence of SCC among males is mainly attributed to a smoking predominance in this gender. On the other hand, ADC is often described as the more frequent LC subtype in females, which may be correlated with a generation of females with a lower smoking history (Gironés et al. 2015; Kabir et al. 2008). Consistently, in the TCGA dataset, females show a two times higher incidence of ADC (54%) than SCC (26%), a pattern common to both in Europeans (42% *vs* 19%) and Africans (6% *vs* 3%), in spite of their quite dissimilar representativeness in TCGA.

Table 3.2. Distribution of ADC and SCC cases per patient gender and ancestry.

Ancestry	ADC Frequency		SCC Frequency	
	MALE	FEMALE	MALE	FEMALE
European	0.34	0.42	0.51	0.19
African-American	0.04	0.06	0.03	0.03
Asian	0.01	0.01	0.01	0.01
Native-American	0.00	0.00	0.00	0.00
Non-Available	0.07	0.05	0.19	0.04
All	0.46	0.54	0.74	0.26

In Table 3.3, the correlation between smoking history expressed as pack per year (PPY) and ADC and SCC subtypes can be evaluated from the TCGA dataset. Here, several significant differences were observed independently of the population group analyzed, with SCC patients

showing always higher smoking loads than ADC (for full TCGA dataset 52.9 PPY in SCC versus 41.8 in ADC; *p*-value < 0.0001). Once again, the TCGA findings are in agreement with previous reports for a strong association of SCC subtype with heavier cigarette smoking histories (The American Cancer Society 2016; Meza et al. 2015).

Interestingly, TCGA data also show a clear pattern for a correlation of ADC with lower smoking load in Africans than in Europeans (30.7 *vs* 43.9 in European; *p*-value 0.0192; Table 3.3). Indeed, several health surveys carried out in the United States have demonstrate that African-Americans are in general lighter smokers (<15 cigarettes per day) than European-Americans, having a preference for menthol cigarettes that are known to release additional cariogenic agents and to contain higher tar levels when compared to conventional cigarettes (Stewart IV 2001). This hypothesis, if true, could explain as well the lower age of ADC onset in Africans when compared with Europeans (60.1 *vs* 65.9; *p*-value 0.0002; Table 3.3). However, the hypothesis of a differentiated effect of menthol cigarettes in LC is not consensual and other environmental and sociocultural factors might play a role as well in the increased susceptibility to ADC by Africans subjects that deserve to be further investigated.

Table 3.3. Smoking history (PPY) and age of onset of ADC and SCC cases in each ancestry.

Ancestry	Smoking history (PPY)		Age at initial diagnosis	
	ADC* (mean)	SCC• (mean)	ADC# (mean)	SCC+ (mean)
European	43.9[a,b]	53.1[a]	65.9[b]	61.1
African-American	30.7[a,b]	52.5[a]	60.1[a,b]	68.3[a]
All	41.8[a]	52.9[a]	66.4	67.2

a - Statistical significant result in the comparison between ADC and SCC cases (P < 0.05).
b - Statistical significant result in comparison within ADC or SCC cases (P < 0.05).
* - Smoking history data provided for 503 ADC patients.
• - Smoking history data provided for 495 SCC patients.
- Age of onset provided for 356 ADC patients.
+ - Age of onset provided for 427 SCC patients.

In addition, we used TCGA dataset to analyze LC anatomic distribution given that ADC and SCC tumors are reported to be situated in distinct lung regions (Table 3.4). While the ADC subtype is more commonly found in peripheral lung sectors, SCC tend to occur more often in proximal areas, closer to airways (The American Cancer Society 2016). However, we could not detect any difference between ADC and SCC regarding their distribution in lung (upper, middle and lower; and right or

left). Here, the lack of statistical significant results can be attributed to a poor resolution of the current lung segmentation concerning the required proximal and distal lung regions. Still, bronchial tumors were identified exclusively in the SCC subtype as it could be predicted based in SCC preferred proximal localization.

Table 3.4. Tumor distribution per lung anatomic site for ADC and SCC cases.

Tumor Anatomical Site	ADC Frequency (N)	SCC Frequency (N)
Upper Lung (Left/Right)	0.590 (308)	0.538 (271)
Middle Lung	0.040 (21)	0.038 (18)
Lower Lung (Left/Right)	0.341 (178)	0.369 (186)
Bronchial	0	0.020 (10)
Other/Non-Available	0.029 (15)	0.038 (19)

N- Absolute number of cases

We anticipated the usage of TCGA database to address the extent of COPD comorbidity in ADC and SCC subtypes. Nevertheless, in both instances limited information regarding lung functional tests (FEV1 and FVC) is provided by TCGA "clinical data" files. Exactly, only 111 ADC and 79 SCC cases, respectively, displayed values for such tests, which prevent the evaluation of COPD prevalence and severity in around 80% of LC cases (Table 3.5).

Table 3.5. COPD Stages of ADC and SCC cases, according to GOLD guidelines.

COPD Stages (GOLD)	ADC Frequency (N)	SCC Frequency (N)
Mild COPD	0.019 (10)	0.024 (12)
Moderate COPD	0.021 (11)	0.040 (20)
Severe COPD	0.013 (7)	0.004 (2)
Very Severe COPD	0.006 (3)	0.006 (3)
No COPD	0.153 (80)	0.083 (42)
Non-Available	0.787 (411)	0.843 (425)

N- Absolute number of cases

3.1.2. Proteolysis related genes analyses

3.1.2.1. Somatic and germline mutations rates

In a first step, to address a potential effect in the ECM remodeling of candidate genes somatic mutations, we examined the ADC and SCC data available through cBioPortal repository. This analysis uncovered several metalloproteases (*ADAM19*, *ADAMTS1*, *MMP2*), collagens (*COL3A1*, *COL4A2*, *COL5A2*, *COL6A3*, *COL14A1*), laminins (*LAMA3*, *LAMB1*), as the top mutated genes with the highest percentages of somatic mutations in both ADC and SCC subtypes (Fig. 3.2 and 3.3). However, those values only account for the total number of mutations per total number of patients, without considering gene size (coding region). Theoretically, if we assume a constant mutation rate across the whole genome, the longer the gene is, the highest probability it has to accumulate a larger number of mutations. Consistently, *COLs* and *LAMAs* identified through cBioPortal webtool were indeed the largest genes included in our candidate series ranging from 438 bp to 11 kbp. Therefore, to obtain a more accurate estimate of each gene mutability, we decided to normalize mutation data by gene size (coding region and neighboring splice sites).

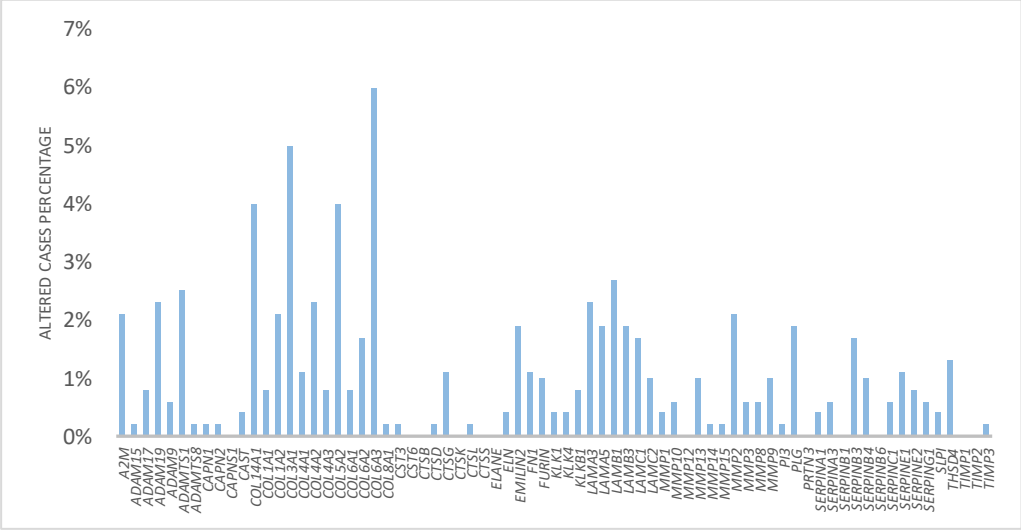

Figure 3.2. Mutation rates as retrieved by cBioPortal for ADC patients. Mutation rates are calculated as total number of somatic mutation per gene, per total number of ADC cases (http://www.cbioportal.org/).

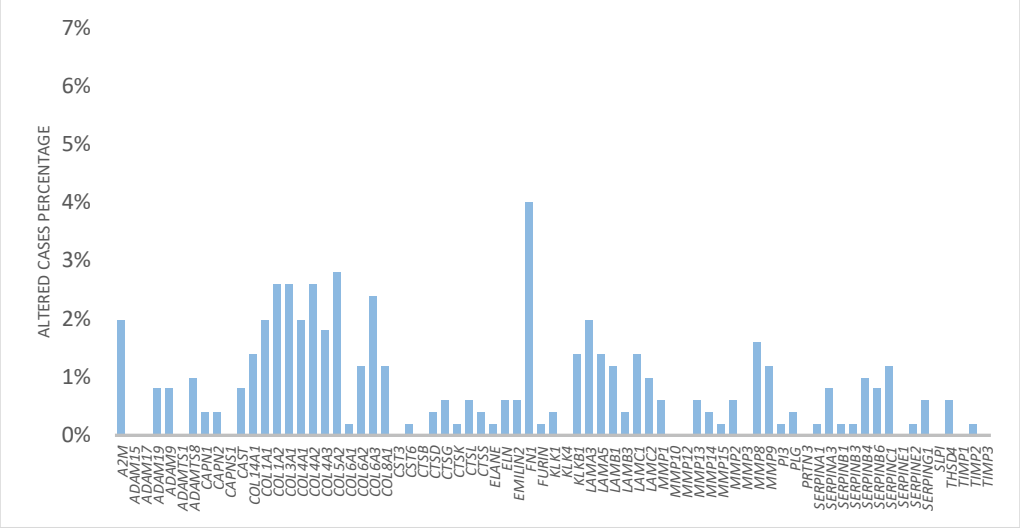

Figure 3.3. Mutation rates as retrieved by cBioPortal for SCC patients. Mutation rates are calculated as total number of somatic mutation per gene, per total number of SCCC cases (http://www.cbioportal.org/).

In addition, to obtain a better appraisal of each gene mutation rate and potential effects in ECM remodeling, tumor microenvironment and LC progression (neutral or non-neutral mutations), we used Polyphen algorithm to predict the functional impact of each mutation. In Figure 3.4 we show the results of normalized mutation rates per each candidate gene and the relative contribution of each functional mutation (benign; possibly; and probably damaging) type using Polyphen predictions. In our analysis, we decided to merge *SERPINB3* and *SERPINB4* mutation data because of the high sequence similarity of these genes (92%), which is expected to interfere with the reads alignments resulting from the deep sequencing methods used by the TCGA project.

Overall, in our analysis for ADC subtype *CTSG* and *MMP2* proteases, *COL3A1/5A2* substrates and *SERPINB3/B4* inhibitors are, not only the genes presenting the highest mutation rates, but also the ones with a larger contribution of somatic mutations with more serious effects in gene function (classified by polyphen as probably damaging, Fig. 3.4).

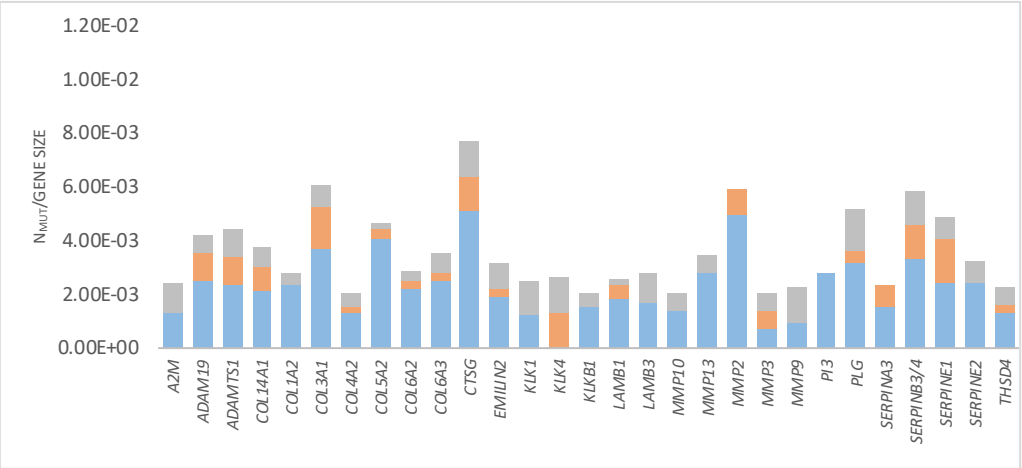

Figure 3.4. Top mutated candidate genes in ADC subtype and mutation functional predictions by Polyphen. Mutation rates were normalized by gene size (coding sequencing and neighboring splice sites). Polyphen predictions as probably damaging mutations are shown in blue (•), as possibly damaging in orange (•), and benign in grey (•). A figure with the full set of candidate genes is available through *Annex (figure A1)*.

On the other hand, SCC subtype displays a distinct mutational pattern from ADC, with *ADAM19*, *CTSG* and *MMP8* proteases, *COL8A1* and *COL1A2* substrates and *PI3* and *TIMP2* inhibitors emerging as the top mutated genes, with a larger fraction of somatic mutations with potential negative effects in the function of these molecules (Fig. 3.5). Altogether, these findings suggest that processes of ECM degradation and remodeling are likely to be affected in different ways in ADC and SCC. Particularly, in ADC somatic mutations seem to be somehow correlated by functional pathways and contribute to the loss fibrillar collagens with functions in lung tensile strength (*COL3A1/A5*), to the inactivation of proteases with a wide spectrum of collagenase (*MMP2*) and chymotrypsin-like (*CTSG*) activities, and to the dysregulation of *SERPINB3/B4* inhibitors controlling *CTSG* and other serine as cysteine proteases. Furthermore, some of these mutations were found to co-occur in the same tumor in an odd association that may not be explained by chance, this is the case of *MMP2/SERPINB3* and *CTSG/SERPINB4*, pairs (Table 3.6).

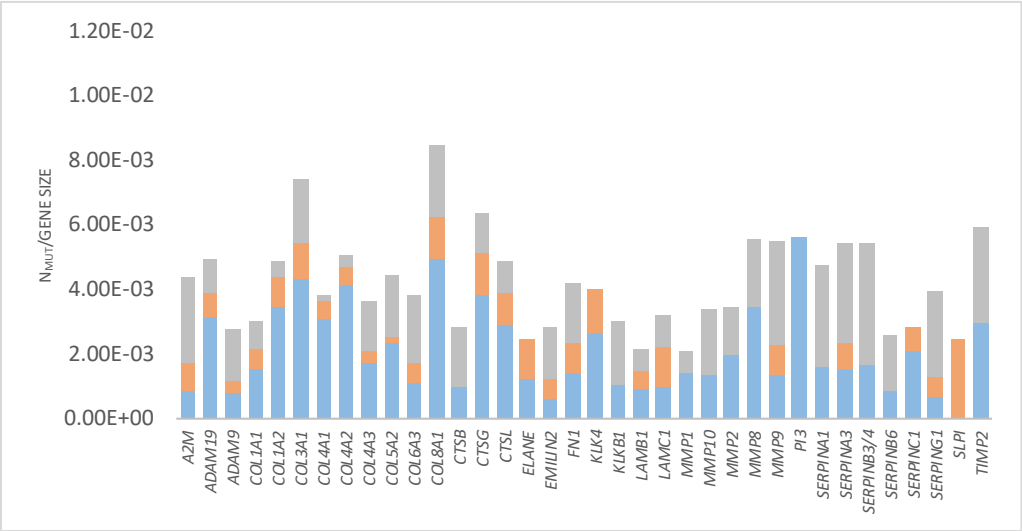

Figure 3.5. Top mutated candidate genes in SCC subtype and mutation functional predictions by Polyphen. Mutation rates were normalized by gene size (coding sequencing and neighboring splice sites). Polyphen predictions as probably damaging mutations are shown in blue (●), as possibly damaging in orange (●), and benign in grey (●). A figure with the full set of candidate genes is available through *Annex (figure A2)*.

Table 3.6. Significant tendency to co-occurrence of candidate genes in ADC cases.

Gene A	Gene B	*p*-Value
MMP2	*SERPINB3*	<0.001
CTSG	*SERPINB4*	0.001
COL5A1	*SERPINB4*	0.011
COL5A1	*CTSG*	0.016
COL3A1	*CTSG*	0.027
COL3A1	*COL5A1*	0.044

Conversely, in SCC, both fibrillar and short chain collagens appear to be affected (*COL3A1*, *COL8A1*) and metalloprotease activity to be compromised by the loss of these enzymes (*ADAM19*, *MMP8*) but also their inhibitors (*TIMP2*). For SCC, evidence of mutation co-occurrence was only detected in *COL8A1/TIMP2* and *CTSG/MMP8* pairs (Table 3.7). Interestingly, *CTSG* identified as one of top mutated genes in ADC and SCC, was also previously identified as a potential protein biomarker for LC, since it was found to be downregulated in patients with LC when compared to non-cancer ones in a high-throughput proteomics screening of bronchoalveolar lavage samples (Carvalho

et al. 2017). Therefore, our findings point to somatic mutations as a probable mechanism for CTSG downregulation in LC.

Table 3.7. Significant tendency to co-occurrence of candidate genes in SCC cases.

Gene A	Gene B	*p*-Value
COL8A1	*TIMP2*	0.012
CTSG	*MMP8*	0.047

The complete evaluation of the somatic mutation landscape for the 73 candidate genes in LC reveals quite divergent mutability patterns in ADC and SCC, with potential implication in ECM assemblage, tumor microenvironment and disease progression. In this context, it is attractive to conjecture if some of these mutations could represent any type of adaptation to opposite preferential localization of ADC in peripheral lung regions and SCC closer to the airways.

In contrary, the analysis of germline variability (MAF<5%) for both ADC and SCC (Fig. 3.6; Fig. 3.7), shows limited differentiation between LC subtypes, where few genes seem to escape from a more common mutability pattern, as it is the case of *CTSB* with a mutation rate three times higher in SCC than ADC. This finding may suggest that *CTSB* variation may play a role in the genetic susceptibility to the SCC subtype. However, as our results may suggest a negative effect due to gene loss of function, *CTSB* has been correlated to LC through gain of function, as it was found overexpressed in aberrant metastasis and lower survival rates (Gong et al. 2013).

These results are also corroborated by analysis with SIFT and CADD scores (*Annex fig. A5, A6, A7, A8, A9, A10, A11, A12*), with some punctual deviations, such as it happens with *SERPINB3/B4* in ADC where predicted number of mutations with negative effects is reduced.

Importantly, if in one hand the similar germline distributions obtained for ADC and SCC point to a low impact of candidate gene variability into LC genetic susceptibility, on the other, these strengthen the hypothesis of a link of the somatic landscape to LC, since it cannot be attributed simply to specific mutability features of each gene.

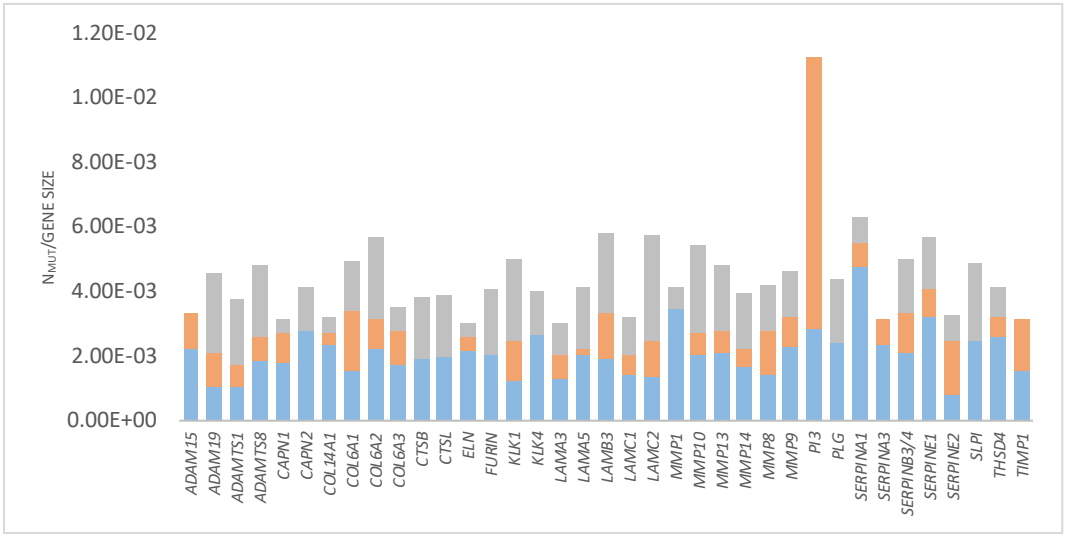

Figure 3.6. Top germline mutated candidate genes in ADC subtype and mutation functional predictions by Polyphen. Mutation rates were normalized by gene size (coding sequencing and neighboring splice sites), with a cutoff of MAF <0.05 of non-Finnish European population from ExAc, excluding common variants. Polyphen predictions as probably damaging mutations are shown in blue (●), as possibly damaging in orange (●), and benign in grey (●). A figure with the full set of candidate genes is available through *Annex (figure A3)*.

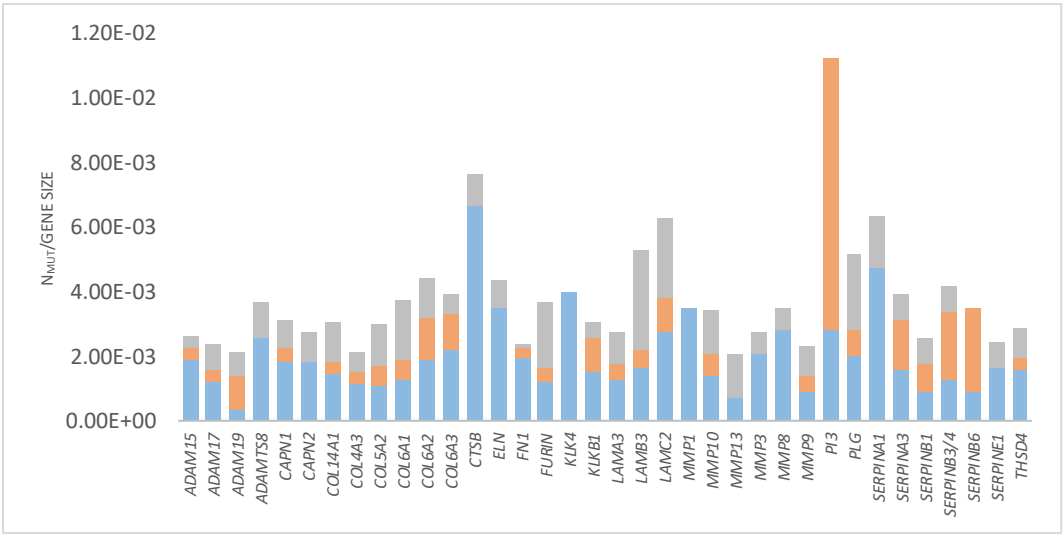

Figure 3.7. Top germline mutated candidate genes in SCC subtype and mutation functional predictions by Polyphen. Mutation rates were normalized by gene size (coding sequencing and neighboring splice sites), with a cutoff of MAF <0.05 of non-Finnish European population from ExAc, excluding common variants. Polyphen predictions as probably damaging mutations are shown in blue (●), as possibly damaging in orange (●), and benign in grey (●). A figure with the full set of candidate genes is available through *Annex (figure A4)*.

3.1.2.2. Candidate gene expression

The analysis of candidate gene expression differences between tumor and non-tumor sections is represented as fold change of transcriptional levels, as shown in Figure 3.10. There, we can notice that most genes evaluated in ADC and SCC subtypes exhibit quite similar expression trends. Generally, genes with upregulated expression in ADC also display augmented levels in SCC, and vice versa for downregulated genes. Still, there are some exceptions like *PI3*, found to be upregulated in SCC cases and slightly downregulated in ADC. Interestingly, *PI3* encodes a small serine protease inhibitor (elafin) with affinity to neutrophil elastase (ELANE) that has been previously reported to be dysregulated in squamous cell cancer types in lung and esophagus, and proposed as a potential target for cancer therapeutics (Yoshida et al. 2002). Oddly, elafin shows an increase expression in highly differentiated tumors in contrast to undifferentiated ones (Yamamoto et al. 1997). More precisely, in breast cancer elafin downregulation has been correlated with an augmented activity of elastase, higher tumor proliferation and shorter times to relapse (Hunt et al. 2013).

This analysis unveiled as well an upregulation of most metalloproteases of the *MMPs* family in both LC subtypes, whereas members of *ADAMTS* family, in spite of sharing a similar activity to *MMPs*, were found to be downregulated. However, ADAMTS1 and -8, have been reported as inhibitors of angiogenesis and therefore potential suppressors of tumor growth (Kumar et al. 2012). The upregulated genes: *MMP1*, *-9*, *-10*, *-12*, and *-13;* belong to distinct MMP clades with diverse functional activities (collagenases, gelatinases and stromelysins), thus having a potential broad range of effects in ECM and lung structural scaffolding, due to their distinct capabilities to cleave elastin and diverse collagen types (I, III, and IV). Among these, MMP1 has already been proposed as a possible biomarker for LC diagnosis and as a treatable target, because of its effects on malignant tumor cells progression if combined with the activity of transcription factor *STAT3* (signal transducer and activator of transcription 3) (Schütz et al. 2015). Moreover, MMP1 overexpression has been associated in other cancer types with poor prognoses, and both MMP12 and -13 expression was described to be altered by malignant cells in diverse squamous cell cancers types, having a potential usage as cancer biomarkers (Reunanen & Kähäri 2013). Also, *MMP9* has been reported to be upregulated in NSCLC, particularly, in larger tumors and metastatic regions, therefore, being also correlated with cancer later stages (El-badrawy et al. 2014). In previous studies, *MMP14* has also been found upregulated in NSCLC cases, although researchers could not evaluate whether MMP14 was effectively proteolytic active. Still, MMP14 was also proposed as potential biomarker and therapeutic target for LC (Atkinson et al. 2007). In LC, the overexpression of several MMPs is likely to contribute to the dysfunction of ECM turnover with expected outcomes in tumor cells proliferation

through ECM stiffening, lung fibrosis and other changes in tissue/cancer microenvironment. Moreover, MMPs increased expression can be as well an indicator of an inflammatory state connected with chronic inflammation and enhanced lung cancer risk (Bonnans et al. 2014).

Remarkably, *MMP2* and *MMP8* found to be among the genes with higher rates of somatic mutations, show only minor differences in gene expression suggesting quite distinct mechanisms of action from the remaining family in LC progression. Notably, *CTSG* previously found to display one of the highest rates of somatic mutations in ADC and SCC was identified to be downregulated in both LC subtypes, point to a possible control of protease activity in LC by two independent mechanisms, by mutation and by gene expression. Altogether, these results are in agreement with previous evidence for lower levels of CTSG in BALF of LC patients and on a possible application of this molecule as a LC biomarker (Carvalho et al 2017).

Although most protease inhibitors were found to be downregulated in LC, *SERPINB3/B4* genes were among the most upregulated, especially in SCC. Oddly, these were one of top somatic mutated genes in ADC suggesting distinct effects of SERPINB3/4 activities in these LC subtypes. One of the most important findings for these genes in cancer research is attributed to the fact that they have been proposed as potential molecular targets for immunotherapy in melanoma. Basically, when *SERPINB3/4* are somatically mutated, the patient survival after treatment improves significantly. Even thought, the mechanisms underlying a better response to melanoma treatment remain unknown, a role connected to SERPINB3/4 innate immunity functions has been advanced (Riaz et al. 2016). Additionally, SERPINB3/4 are usually secreted by NSCLC tumors, and are thought to be key factors in aberrant epithelial proliferation (Calabrese et al. 2012), increasing their importance as possible LC biomarkers.

Concerning ECM substrates, most genes analyzed were found to be downregulated in tumor sections, except for *COL1A1/3A1/5A2*, some of them previously found to be top mutated genes in ADC and/or SCC (Fig. 3.4 and 3.5). This finding may suggest that these genes in general are upregulated to increase ECM deposition and fibrosis in tumor surroundings, but in some subjects, carrying somatic mutations in these genes, the processes of ECM remodeling can be further modified. Actually, collagen type I deposition (*COL1Ax* genes) has been correlated with tumor proliferation and changes in microenvironment caused by abnormal ECM stiffness (Fang et al. 2014; Shintani et al. 2008), which contribute to cancer isolation and protection against natural immune response to cancer cells. The aberrant production of chains of procollagen types I and III (*COL1Ax*, *COL3Ay* genes) has been described in other cancer types, such as in ovarian tumor tissue, (Kauppila et al. 1996). Also, in mouse lung after chemical carcinogenesis, collagen type V (*COL5Ax*) in low quantities has also been related with decreased apoptosis, with a possible contribution to tumor

growth. This is facilitated by a reduction in the co-polymerization of type V with types I and III, which in turn can trigger cancer cell invasion and motility (Parra et al. 2010). Nevertheless, more recent studies have demonstrated a positive influence of type III collagen in the microenvironment regulation, by maintaining normal stromal organization, and by reducing lung metastasis in other types of cancer, like breast cancer (Brisson et al. 2015).

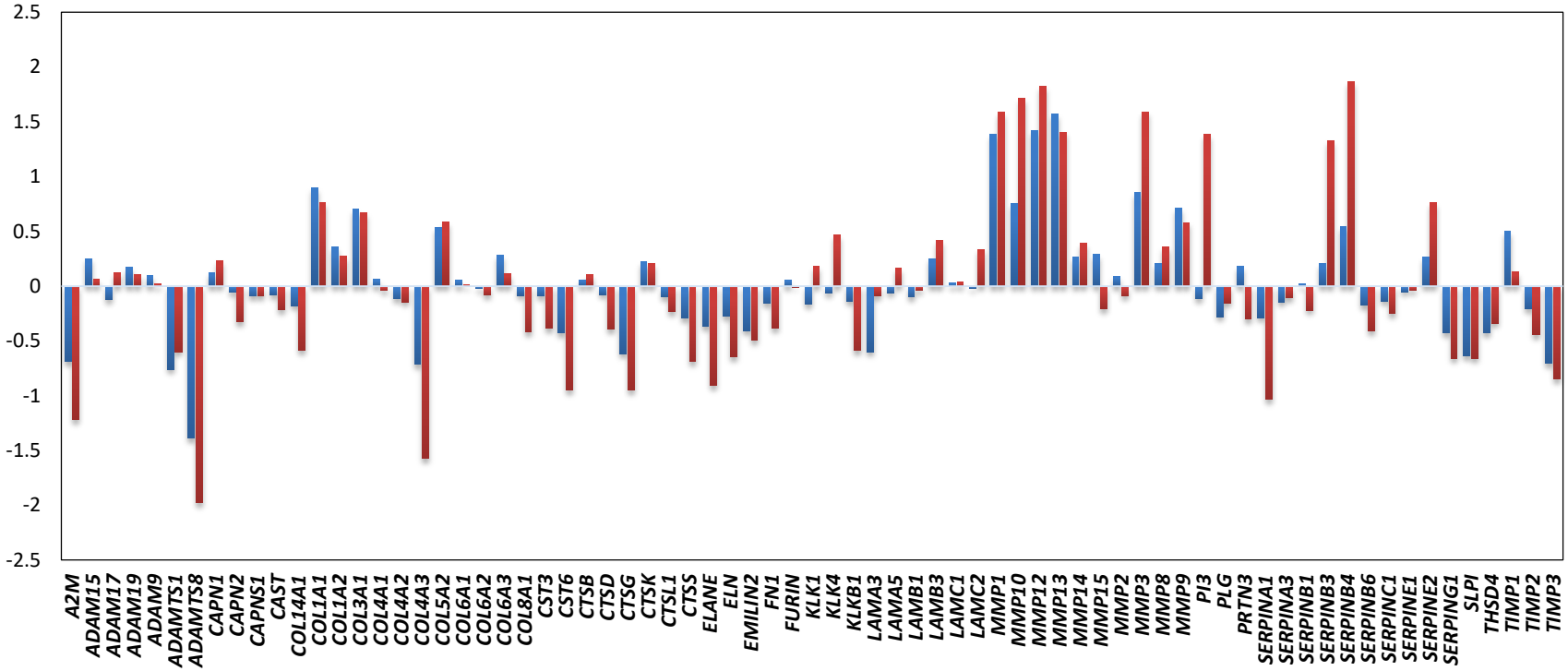

Figure 3.8. Expression change of candidate genes in normal and tumor tissue, with normalization by a logarithmic scale of fold change. Genes with expression above 0 present an upregulation in tumor tissue when compared to normal, while bellow 0 are downregulated in tumor sites. In blue are ADC cases, and in red SCC ones.

3.2. Screening of *SERPINB3*, *SERPINB4* and *CTSG* genes in Portuguese COPD and LC cases

We selected *SERPINB3*, *SERPINB4* and *CTSG* genes for a follow up study, in which sequence variability was surveyed in both COPD and LC cases. We based our candidate gene selection in the previous analysis of TCGA database and literature, considering *SERPINB3/4* as potential targets for cancer treatment (melanoma and eventually LC) and CSTG as protease regulated by SERPINB4, whose gene was also found to be downregulated in LC (Carvalho et al. 2017; Riaz et al. 2016).

In our study, we completed the sequencing analysis of *SERPINB3*/*B4* (exons I-VIII) for only 33 out of the 43 COPD cases collected. Concerning LC samples due to their low DNA quality, we were only capable of covering 5 in 8 exons (I, IV, VI, VIII) in 21 cases. The screening of *CTSG* included 3 out of 5 exons in 14 COPD cases, and 2 exons in 23 LC cases.

In *SERPINB3*, we identified a total of 13 variants, in which 6 were low-frequency variants (MAF <5%) according with the data available from the 1000 Genomes project for the IBS sample (our control group) (Table 3.8). Remarkably, most of *SERPINB3* variants were detected exclusively in COPD samples (12 out of 13 variants). *SERPINB3* variants with potential functional consequences includes: 1) Three variants located in *SERPINB3* 5'UTR (c.-124G>A, c.-97C>T and c.-65T>C), in a genome sequence spanning several binding sites for transcript factors as identified by ENCODE chip-seq studies (Figure 3.11); 2) Three nonsynonymous mutations (p.Val134Ile, p.Asp240Tyr and p.Trp269Arg) with deleterious prediction by SIFT and/or Polyphen; 3) Two common nonsynonymous substitutions (p.Gly351Ala and p.Thr357Ala) located in SERPINB3 RCL in close proximity to the scissile bound (P1-P1'; Figure 3.12), which were reported to increase the susceptibility to other diseases. In fact, p.Gly351Ala has been linked to liver cirrhosis, thus suggesting a possible influence of this variant in fibrosis development (Turato et al. 2009). Oddly, this variant has also been reported by *Riaz et al* (2016) as one of the variants probably contributing to a better response to melanoma treatment. Nevertheless, the identification of p.Gly351Ala in our COPD and LC samples, as well as, in previous studies indicates that this is more likely to be a germline mutation rather than a somatic. Furthermore, p.Gly351Ala substitution has been described as affecting SERPINB3 ability to counteract several parasite-derived cysteine proteases. (Kanaji et al. 2007) and RCL hydrophobicity has also been reported to be an important factor for its binding to hepatitis B virus (Chen et al. 2005). There, the placement of a variant such as p.Gly351Ala may hint a possible an impact virus binding to the RCL of SERPINB3.

In the comparison of p.Thr357Ala frequencies between COPD patients and controls, we found a Fisher's exact test significant *p*-value (0.0368). This variant was found to be increased in COPD and in a lower extent in LC samples, thus possibly conferring increased susceptibility to lung disease.

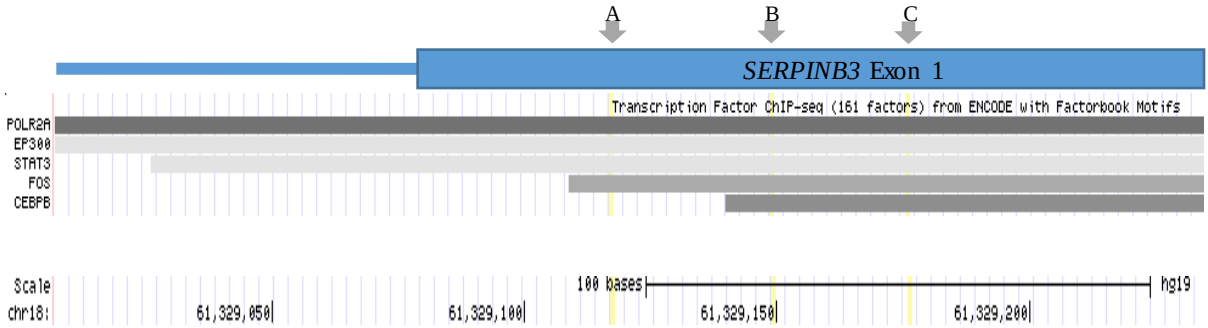

Figure 3.9. Schematic representation of *SERPINB3* 5'UTR variants location. Coincidence of c.-65T>C (A), c.-97C>T (B), and c.-124G>A (C) variants in *SERPINB3* 5'UTR with transcription factors binding sites, including POLR2A, EP300, STAT3, FOS, and CEBPB. The inset shows the transcription factors binding region with the locations of variants retrieved within the UCSC Genome Browser view.

Figure 3.10. *SERPINB3* protein structure with detected variants positions in reactive center loop highlighted. *SERPINB3* locations of variants in the reactive center site loop detected by Sanger sequencing in a Portuguese cohort, indicated by grey arrows, Gly351 and Thr357.

Table 3.8. Variants identified in our cohort of Portuguese COPD and LC patients.

GENE	SNP	Consequence	MAF			Bioinformatics Prediction	
			COPD Cases	LC Cases	Controls - 1000 Genomes IBS	SIFT scores	PolyPhen scores
	rs470750	c.-124G>A	0.0454 (N=33)	0 (N=21)	0	NA	NA
	rs471018	c.-97C>T	0.0303 (N=33)	0 (N=21)	0	NA	NA
	rs471017	c.-65T>C	0.0152 (N=33)	0 (N=21)	0	NA	NA
	rs781685229	p.Asp14Asp	0 (N=33)	0.05 (N=10)	0	NA	NA
	rs61754491	p.Val134Ile	0.0303 (N=33)	0 (N=21)	0.014	Deleterious (0.03)	Benign (0.062)
	rs148254791	p.Val134Ala	0.0303 (N=33)	0 (N=21)	0.014	Tolerated (0.22)	Benign (0.052)
SERPINB3[a]	rs61733410	p.Asp240Tyr	0.0152 (N=33)	NS	0.0047	Deleterious (0)	Probably damaging (0.945)
	rs150069636	p.Trp269Arg	0.0303 (N=33)	0 (N=21)	0.0467	Deleterious (0)	Probably damaging (0.998)
	rs3180227	p.Gly351Ala	0.197 (N=33)	0.1428 (N=21)	0.1729	Tolerated (0.83)	Benign (0.001)
	rs1065205	p.Thr357Ala	0.1364* (N=33)	0.119 (N=21)	0.0561*	Tolerated (0.81)	Benign (0)
	rs12953909	p.Glu362Lys	0.0303 (N=33)	0 (N=21)	0.0047	Tolerated (0.49)	Benign (0.061)
	rs61703421	c.*42C>G	0.1515 (N=33)	0.1428 (N=21)	0.1822	NA	NA
	NA	c.*190A>G	0.0303 (N=33)	0 (N=21)	NS	NA	NA
SERPINB4[b]	rs55883719	p.Asn25Ser	0.0119 (N=42)	NS	0.0093	Tolerated (0.12)	Benign (0.085)
CTSG[c]	rs45567233	p.Asn125Ser	0 (N=14)	0.1087 (N=23)	0.0514	Tolerated (0.40)	Benign (0)

NA - Not applicable.
NS - Region not surveyed.
* - Significant P-value with Fisher test.
[a] – The ENSP00000283752.5 transcript was used in *SERPINB3* mutation annotation.
[b] – The ENSP00000343445.5 transcript was used in *SERPINB4* mutation annotation.
[c] – The ENSP00000216336.2 transcript was used in *CTSG* mutation annotation.

Taking into account a possible enrichment of low frequency variants (MAF <0.05) in COPD cases, we applied Burden and C-alpha tests to evaluate the statistical significance of these findings. We considered only *SERPINB3* nonsynonymous and UTR variants, separately and combined, and COPD and LC groups

alone, and merged all together in a single lung disease group. As control, we used 1000 Genomes data from the IBS sample, where only variants located in the region of our sequencing screening were considered. No statistical significant results were observed between these groups (Table 3.9). Nevertheless, in the comparison of COPD cases against controls, for *SERPINB3* UTR variants only, and for all *SERPINB3* variants, *p*-values close to the 0.05 significant threshold were obtained. This interesting trend needs to be confirmed by an extended study with larger sample sizes. Only with the accumulation of additional low-frequency variants this result may be reach statistical significance.

Table 3.9. Statistical tests for low frequency variants found in *SERPINB3* and *SERPINB4* genes.

Tests	Burden test *p*-Value	C-alpha *p*-Value
***SERPINB3* missense COPD vs Controls**	0.2663	0.6211
***SERPINB3* UTR COPD vs Controls**	0.0650	0.0551
***SERPINB3* missense LC vs Controls**	1.0000	0.2799
***SERPINB3* UTR LC vs Controls**	1.0000	1.0000
***SERPINB3* missense COPD vs LC**	1.0000	0.3480
***SERPINB3* UTR COPD vs LC**	1.0000	0.5139
***SERPINB4* missense COPD vs Controls**	1.0000	1.0000
***SERPINB3* All vs Controls**	0.4156	0.1272
***SERPINB3* All COPD vs Controls**	0.0669	0.1125
***SERPINB3* All LC vs Controls**	1.000	1.000

Given the current understanding of SERPINB3 activity, and if proven the current variation trends in COPD, these are more likely to be correlated with immune homeostasis, modulation of the inflammatory response and some levels of lung fibrosis in spite of this not being a common finding in COPD (Turato & Pontisso 2015). Moreover, SERPINB3 may be important in COPD development, since its expression is augmented in patients after cigarette smoking, while in other subjects remains unchanged, which can be attributed to an upregulation of this protein in altered COPD bronchial tree (Franciosi et al. 2014).

For *SERPINB4*, we found a single variant in COPD cases only and, in contrary, for *CTSG* we found a unique variant in LC patients. No significant test of association to lung disease was obtained for these genes (Table 3.9).

A limitation of this study is due to the low sample size of our cohorts and the lack of a Portuguese control group. On the other hand, the usage of Sanger sequencing methods may have prevented the detection of somatic mutations present in COPD and LC cases. Since almost all detected variants had been identified somewhere else in control databases such as ExAc, pointing out that these are most likely to be germline

mutations. In order to identify somatic and germline mutations, we should have used deep-sequencing methods with considerable large coverages (>30x).

3.3. Concluding remarks

The analysis of the clinical and epidemiological data of LC cases (ADC and SCC subtypes) provided by TCGA, in general agrees with previous reports, namely in the considerable incidence of LC (ADC) in African-Americans, at younger ages and in spite of their significant lower smoking load. This results highlight a need to pursue further studies in this population group to better understand the causes for an increased LC susceptibility in African-Americans. In addition, this dataset also supports a stronger association of SCC to heavier smoking, mainly observed in males. Conversely, this dataset failed to provide an evaluation of COPD comorbidity in LC, due to the lack of annotation for lung function measurements (FEV1/FVC) in the vast majority of ADC and SCC cases.

In the evaluation of the mutation rates for the 73 proteolysis genes using the TCGA dataset, we found that ADC and SCC subtypes are largely divergent regarding their somatic landscape, but quite similar in their germline patterns. Overall, this analysis allowed the identification of several highly mutated genes like *CTSG*, *SERPINB3/B4*, *MMP2*, *COL3A1*, and *COL5A2*, in which the loss of function may contribute to modify ECM assemblage according to LC localization, in more proximal regions for the SCC subtype and in more distal sector in ADC. The analysis of candidate gene expression showed some concordant trends with the distribution of somatic mutation, such as in the case of *CTSG* where both mechanisms appear to contribute to a gene downregulation in LC. While in other circumstances discordant results were obtain, like in the situation of *COL3A1/5A2* that may be correlated with LC heterogeneity across different patients.

In our study of Portuguese COPD and LC patients, we collected preliminary evidence for a role of *SERPINB3* in COPD susceptibility. We detected not only an increased frequency in COPD cases of a variant (p.Thr357Ala) localized in the RCL, probably affecting SERPINB3 activity, as well as a potential enrichment in COPD patients of variants with possible functional consequences.

4. References

Alberg, A.J., Brock, M. V. & Samet, J.M., 2005. Epidemiology of lung cancer: Looking to the future. *Journal of Clinical Oncology*, 23(14), pp.3175–3185.

Alder, J.K. et al., 2011. Telomere length is a determinant of emphysema susceptibility. *American Journal of Respiratory and Critical Care Medicine*, 184(8), pp.904–912.

Annoni, R. et al., 2012. Extracellular matrix composition in COPD. *European Respiratory Journal*, 40(6), pp.1362–1373.

Arpino, V., Brock, M. & Gill, S.E., 2015. The role of TIMPs in regulation of extracellular matrix proteolysis. *Matrix Biology*, 44–46, pp.247–254.

Askew, D.J. & Silverman, G.A., 2008. Intracellular and extracellular serpins modulate lung disease. *Journal of Perinatology*, pp.127–135.

Atkinson, J.M. et al., 2007. Membrane type matrix metalloproteinases (MMPs) show differential expression in non-small cell lung cancer (NSCLC) compared to normal lung: Correlation of MMP-14 mRNA expression and proteolytic activity. *European Journal of Cancer*, 43(11), pp.1764–1771.

Auton, A. et al., 2015. A global reference for human genetic variation. *Nature*, 526(7571), pp.68–74.

Bachman, H. et al., 2015. Utilizing Fibronectin Integrin-Binding Specificity to Control Cellular Responses. *Advances in Wound Care*, 4(8), pp.501–511.

Balestrini, J.L. et al., 2016. Extracellular matrix as a driver for lung regeneration. *Annals of Biomedical Engineering*, 43(3), pp.568–576.

Bidan, C.M. et al., 2015. Airway and Extracellular Matrix Mechanics in COPD. *Frontiers in Physiology*, 6, pp.346.

Bonnans, C., Chou, J. & Werb, Z., 2014. Remodelling the extracellular matrix in development and disease. *Nature reviews molecular cell biology*, 15(12), pp.786–801.

Bowler, R.P., Barnes, P.J. & Crapo, J.D., 2004. The Role of Oxidative Stress in Chronic Obstructive Pulmonary Disease. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 1(2), pp.255–277.

Bracci, P.M. et al., 2012. Cigarette smoking associated with lung adenocarcinoma in situ in a large case-control study (SFBALCS). *Journal of Thoracic Oncology*, 7(9), pp.1352–1360.

Brew, K. & Nagase, H., 2011. The tissue inhibitors of metalloproteinases (TIMPs): An ancient family with structural and functional diversity Keith. *Biochimica Biophysica Acta*, 1803(1), pp.55–71.

Brisson, B.K. et al., 2015. Type III collagen directs stromal organization and limits metastasis in a murine model of breast cancer. *American Journal of Pathology*, 185(5), pp.1471–1486.

Bromme, D. & Wilson, S., 2011. Role of Cysteine Cathepsins in Extracellular Proteolysis. *Extracellular Matrix Degradation*, pp.23–52.

Burgess, J.K. et al., 2016. The Extracellular Matrix - the under-recognised element in lung disease? *The Journal of pathology*, pp.397–409.

Burgstaller, G. et al., 2017. The instructive extracellular matrix of the lung: basic composition and alterations in chronic lung disease. *European Respiratory Journal*, 50(1), p.1601805.

Bustamante Alvarez, J.G. et al., 2015. Advances in immunotherapy for treatment of lung cancer. *Cancer Biology & Medicine*, 12(3), pp.209–222.

Calabrese, F. et al., 2012. Serpin B4 isoform overexpression is associated with aberrant epithelial proliferation and lung cancer in idiopathic pulmonary fibrosis. *Pathology*, 44(3), pp.192–198.

Caley, M.P., Martins, V.L.C. & Toole, E.A.O., 2015. Metalloproteinases and Wound Healing. *Advances in Wound Care*„ 4(4), pp.225–234.

Carvalho, A.S. et al., 2017. Bronchoalveolar Lavage Proteomics in Patients with Suspected Lung Cancer. *Nature Publishing Group*, (January), pp.1–13.

Casado, B. et al., 2007. Protein expression in sputum of smokers and chronic obstructive pulmonary disease patients: A pilot study by CapLC-ESI-Q-TOF. *Journal of Proteome Research*, 6(12), pp.4615–4623.

Cathcart, J. et al., 2015. Targeting Matrix Metalloproteinases in Cancer: Bringing New Life to Old Ideas. Genes & Disease, 2, pp.26–34.

Chang, J.T.-H., Lee, Y.M. & Huang, R.S., 2015. The impact of the Cancer Genome Atlas on lung cancer. *Translational research : the journal of laboratory and clinical medicine*, 166(6), pp.568–85.

Chang, K. et al., 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10), pp.1113–1120.

Chen, M. et al., 2005. Hydrophobicity of reactive site loop of SCCA1 affects its binding to hepatitis B virus. *World Journal of Gastroenterology*, 11(19), pp.2864–2868.

Cho, M.H. et al., 2010. Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nature Genetics*, 42(3), pp.200–202.

Churg, A., Zhou, S. & Wright, J.L., 2012. Matrix metalloproteinases in COPD. *European Respiratory Journal*, 39(1), pp.197–209.

Cibulskis, K. et al, 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3), pp.213-219

Cortinovis, D. et al., 2016. Targeted therapies and immunotherapy in non-small-cell lung cancer. *Ecancermedicalscience*, 10, pp.1–28.

Cox, T.R. & Erler, J.T., 2011. Remodeling and homeostasis of the extracellular matrix : implications for fibrotic diseases and cancer. *Disease Models & Mechanisms*, 178, pp.165–178.

Crosby, L.M. & Waters, C.M., 2010. Epithelial repair mechanisms in the lung. *AJP: Lung Cellular and Molecular Physiology*, 298(6), pp.L715–L731.

Currie, G.P., Kennedy, A.-M. & Denison, A.R., 2009. Tools used in the diagnosis and staging of lung cancer: What's old and what's new? *Qjm*, 102(7), pp.443–448.

Dang, N., Meng, X. & Song, H., 2016. Nicotinic acetylcholine receptors and cancer (Review). *Biomedical Reports*, pp.515–518.

Denden, S. et al., 2010. Alpha-1 antitrypsin gene polymorphism in Chronic Obstructive Pulmonary Disease (COPD). *Genetics and Molecular Biology*, 33(1), pp.23–26.

Domej, W. & Oettl, K., 2014. Oxidative stress and free radicals in COPD – implications and relevance for treatment. *International Journal of Chronic Obstructive Pulmonary Disease*, pp.1207–1224.

Duffy, M.J. et al., 2011. The ADAMs family of proteases: new biomarkers and therapeutic targets for cancer? *Clinical Proteomics*, 8(1), p.9.

Dunsmore, S.E., 2008. Treatment of COPD: A matrix perspective. *International Journal of COPD*, 3(1), pp.113–122.

Dunsmore, S.E., Eugene, D. & Eugene, D., 1996. Extracellular matrix biology in the lung. *American Journal of Physiology - Lung Cellular and Molecular Physiology*, 270(1), L3-27.

Durham, A.L. & Adcock, I.M., 2015. The relationship between COPD and lung cancer. *Lung Cancer*, 90(2), pp.121–127.

Edwards, D.R., Handsley, M.M. & Pennington, C.J., 2009. The ADAM metalloproteinases. *Molecular Aspects of Medicine*, 29(5), pp.258–289.

Eilbeck, K., Quinlan, A. & Yandell, M., 2017. Settling the score: variant prioritization and Mendelian disease. *Nature Reviews Genetics*, 18(10), pp.599–612.

Eisenhut, F. et al., 2017. FAM13A is associated with non-small cell lung cancer (NSCLC) progression and controls tumor cell proliferation and survival. *OncoImmunology*, 6(1), p.e1256526.

El-badrawy, M.K. et al., 2014. Matrix Metalloproteinase-9 Expression in Lung Cancer. *Journal of Bronchology and Interventional Pulmonology*, 21(4), pp.327–334.

Enewold, L. et al., 2012. SERPINA1 and ELA2 polymorphisms are not associated with COPD or lung cancer. *Anticancer research*, 32(9), pp.3923–8.

Eurlings, I.M.J. et al., 2014. Similar matrix alterations in alveolar and small airway walls of COPD patients. *BMC pulmonary medicine*, 14(1), p.90.

Fang, M. et al., 2014. Collagen as a double-edged sword in tumor progression. *Tumor Biology*, 35(4), pp.2871–2882.

Field, J.K. et al., 2013. CT screening for lung cancer: Countdown to implementation. *The Lancet Oncology*, 14(13), pp.591–600.

Fischer, B.M., Pavlisko, E. & Voynow, J.A., 2011. Pathogenic triad in COPD: Oxidative stress, protease-antiprotease imbalance, and inflammation. *International Journal of COPD*, 6(1), pp.413–421.

Fonovic, M. & Turk, B., 2014. Cysteine cathepsins and extracellular matrix degradation. *Biochimica et Biophysica Acta*, 1840, pp.2560–2570.

Fortelny, N. et al., 2014. Network Analyses Reveal Pervasive Functional Regulation Between Proteases in the Human Protease Web. *PLoS Biology*, 12(5), p.e1001869.

Franciosi, L. et al., 2014. Susceptibility to COPD: Differential proteomic profiling after acute smoking. *PLoS ONE*, 9(7), pp.3–11.

Frantz, C. et al., 2010. The extracellular matrix at a glance The Extracellular Matrix at a Glance. *Journal of Cell Science*, 123(24), pp.4195–4200.

Frazer, K.A. et al., 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164), pp.851–861.

Gabrielson, E., 2006. Worldwide trends in lung cancer pathology. *Respirology*, 11(5), pp.533–538.

Gettins, P.G.W. & Olson, S.T., 2016. Inhibitory serpins. New insights into their folding, polymerization, regulation and clearance. *Biochemical Journal*, 473(15), pp.2273–2293.

Giebeler, N. & Zigrino, P., 2016. A Disintegrin and Metalloprotease (ADAM): Disintegrin and Metalloprotease Historical Overview of Their Functions Historical Overview of Their Functions. *Toxins*, 8(4), p.122.

Gironés, R. et al., 2015. Ten years of lung cancer in a single center: gender, histology, stage and survival. *Journal of Cancer Metastasis and Treatment*, 1(3), p.201.

Gong, F. et al., 2013. Cathepsin B as a potential prognostic and therapeutic marker for human lung squamous cell carcinoma. *Molecular Cancer*, 12(1), p.125.

González-Arriaga, P. et al., 2012. Genetic polymorphisms in MMP 2, 9 and 3 genes modify lung cancer risk and survival. *BMC Cancer*, 12(1), p.121.

Gooptu, B. & Lomas, D.A., 2009. Conformational Pathology of the Serpins : Themes , Variations , and Therapeutic Strategies.*Annual Review of Biochemestry*, 78, pp.146-176.

Greenlee, K.J., Werb, Z. & Kheradmand, F., 2007. Matrix metalloproteinases in lung: multiple, multifarious, and multifaceted. *Physiological reviews*, 87(1), pp.69–98..

Hamill, K.J. et al., 2009. Laminin deposition in the extracellular matrix : a complex picture emerges. *Journal of Cell Science*, 122(24), pp.4409-4417

Haq, I. et al., 2010. Association of MMP-2 polymorphisms with severe and very severe COPD: a case control study of MMPs-1, 9 and 12 in a European population. *BMC medical genetics*, 11, p.7.

Hardin, M. et al., 2012. CHRNA3/5, IREB2, and ADCY2 are associated with severe chronic obstructive pulmonary disease in Poland. *American Journal of Respiratory Cell and Molecular Biology*, 47(2), pp.203–208.

Harju, T. et al., 2010. Variability in the precursor proteins of collagen I and III in different stages of COPD. *Respiratory Research*, 11(1), p.165.

Herbst, R., Heymach, J. & Lippman, S., 2008. Lung cancer. *New England Journal of Medicine*, 359(13), pp.1367–1380.

Higgins, R., Lewis, C. & Warren, H., 2003. Lung cancer in African Americans. *Annals of Thoracic Surgery*, 76(4), pp.S1363-6.

Houghton, A.M., 2015. Matrix metalloproteinases in destructive lung disease. *Matrix Biology*, 44–46, pp.167–174.

Houghton, a M., 2013. Mechanistic links between COPD and lung cancer. *Nature reviews. Cancer*, 13(4), pp.233–45.

Humphrey, J.D., Dufresne, E.R. & Schwartz, M.A., 2015. Mechanotransduction and extracellular matrix homeostasis. , 15(12), pp.802–812.

Humphrey, L. et al., 2013. Screening for Lung Cancer With Low-Dose Computed Tomography : A Systematic Review to Update the U.S. Preventive Services Task Force Recommendation. *Annals of Internal Medicine*, 159(6).

Hunt, K.K. et al, 2013. Elafin, an inhibitor of elastase, is a prognostic indicator in breast cancer. *Breast Cancer Research*, 15(1):R3.

Hynes, R.O., 2013. Extracellular matrix: not just pretty fibrils. *Science*, 326(5957), pp.1216–1219.

Hynes, R.O. & Naba, A., 2012. Overview of the Matrisome — An Inventory of Extracellular Matrix Constituents and Functions. *Cold Spring Harbor Perspectives in Biology*, 4(1), pp.1–16.

Ito, K. & Barnes, P.J., 2009. COPD as a disease of accelerated lung aging. *Chest*, 135(1), pp.173–180.

Jackson, V.E. et al., 2016. Exome-wide analysis of rare coding variation identifies novel associations with COPD and airflow limitation in *MOCS3* , *IFIT3* and *SERPINA12*. *Thorax*, 71(6), pp.501-509.

John, A., Mcguinness, A. & Sapey, E., 2017. Oxidative Stress in COPD : Sources , Markers , and Potential Mechanisms. *Journal of Clinical Medicine*, 6(2).

Kabir, Z., Connolly, G.N. & Clancy, L., 2008. Sex-differences in lung cancer cell-types? An epidemiologic study in Ireland. *Ulster Medical Journal*, 77(1), pp.31–35.

Kanaji, S. et al., 2007. Squamous cell carcinoma antigen 1 is an inhibitor of parasite-derived cysteine proteases. *FEBS Letters*, 581(22), pp.4260–4264.

Kasabova, M. et al., 2011. Cysteine cathepsins: Markers and therapy targets in lung disorders. *Clinical Reviews in Bone and Mineral Metabolism*, 9(2), pp.148–161.

Kauppila, S. et al., 1996. Expression of mRNAs for type I and type III procollagens in serous ovarian cystadenomas and cystadenocarcinomas. *American Journal Of Pathology*, 148(2), pp.539–548.

Kelwick, R. et al., 2015. The ADAMTS (A Disintegrin and Metalloproteinase with Thrombospondin motifs) family. *Genome Biology*, 16, pp.113–129.

Khiroya, H. & Turner, A.M., 2015. The role of iron in pulmonary pathology. *Multidisciplinary Respiratory Medicine*, 10, pp.1–7.

Kim, W. & Lee, S., 2015. Candidate genes for COPD : current evidence and research. *International Journal of Chronic Obstructive Pulmonary Disease*, 10, pp.2249–2255.

Kniazeva, E. & Putnam, A.J., 2009. Endothelial cell traction and ECM density influence both capillary morphogenesis and maintenance in 3-D. *American Journal of Physiology. Cell Physiology*, 297(1), pp.179–187.

Korkmaz, B. et al., 2010. Neutrophil Elastase , Proteinase 3 , and Cathepsin G as Therapeutic Targets in Human Diseases. *Pharmacological Reviews*, 62(4), pp.726–759.

Kranenburg, A.R. et al., 2006. Enhanced bronchial expression of extracellular matrix proteins in chronic obstructive pulmonary disease. *American journal of clinical pathology*, 126(5), pp.725–35.

Kugler, M.C. et al., 2015. Sonic hedgehog signaling in the lung: From development to disease. *American Journal of Respiratory Cell and Molecular Biology*, 52(1), pp.1–13.

Kumar, S. et al., 2012. Emerging Roles of ADAMTSs in Angiogenesis and Cancer. *Cancers*, 4(4), pp.1252-1299.

Lazarus, D.R. & Ost, D.E., 2013. How and when to use genetic markers for nonsmall cell lung cancer. *Current opinion in pulmonary medicine*, 19(4), pp.331–9.

Lek, M. et al., 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature Publishing Group*, 536(7616), pp.285–291.

Liou, M.-Y. & Storz, P., 2010. Reactive oxygen species in cancer. *Free Radical Research*, 44(5), pp.479-496.

Löffek, S., Schilling, O. & Franzke, C.W., 2011. Series "matrix metalloproteinases in lung health and disease" edited by J. Müller-Quernheim and O. Eickelberg number 1 in this series: Biological role of matrix metalloproteinases: A critical balance. *European Respiratory Journal*, 38(1), pp.191–208.

Lu, P. et al., 2011. Extracellular matrix degradation and remodeling in development and disease. *Cold Spring Harbor Perspectives in Biology*, 3(12), pp.1–24.

Lucas, S.D. et al., 2011. Targeting COPD : Advances on Low- Molecular-Weight Inhibitors of Human Neutrophil Elastase. *Medicinal Research Reviews*, 33(1), pp.73-101.

Mammoto, A. et al., 2009. A mechanosensitive transcriptional mechanism that controls angiogenesis. *Nature*, 457(7233), pp.1103–1108.

Manolio, T.A. et al., 2010. Finding the missing heritability of complex diseases. *Nature*, 461(7265), pp.747–753.

Mao, Y. & Schwarzbauer, J.E., 2005. Fibronectin fibrillogenesis, a cell-mediated matrix assembly process. *Matrix Biology: Journal of the International Society for Matrix Biology*, 24, pp.389–399.

Marchand, L. Le et al., 2009. Smokers with the CHRNA Lung Cancer-Associated Variants are Exposed to Higher Levels of Nicotine Equivalents and a Carcinogenic Tobacco-Specific Nitrosamine. *Cancer Research*, 68(22), pp.9137–9140.

Matheson, M.C. et al., 2005. Biological dust exposure in the workplace is a risk factor for chronic obstructive pulmonary disease. *Thorax*, 60(8), pp.645–51.

Merid, S.K., Goranskaya, D. & Alexeyenko, A., 2014. Distinguishing between driver and passenger

mutations in individual cancer genomes by network enrichment analysis. *BMC bioinformatics*, 15(1), p.308.

Meza, R. et al., 2015. Lung cancer incidence trends by gender, race and histology in the United States, 1973-2010. *PLoS ONE*, 10(3), pp.1–14.

Mitchell, K.J., 2012. What is complex about complex disorders? *Genome biology*, 13, p.237.

Mithieux, S.M. & Weiss, A.S., 2006. Elastin. *Advances in protein chemistry*, 70(4), pp.437–461.

Mocchegiani, E., Giacconi, R. & Costarelli, L., 2011. Metalloproteases / anti-metalloproteases imbalance in chronic obstructive pulmonary disease : genetic factors and treatment implications. *Current Opinion in Pulmonary Medicine*, 1, pp.11-19.

Moroy, G. et al., 2012. Neutrophil Elastase as a Target in Lung Cancer. *Anti-Cancer Agents in Medicinal Chemistry*, 12(6), pp.565–579.

Mouw, J.K. et al., 2015. Extracellular matrix assembly: a multiscale deconstruction. *Nature Reviews. Molecular Cell Biology*, 15(12), pp.771–785.

Nielsen, A.O. et al., 2017. Variants of the ADRB2 Gene in COPD: Systematic Review and Meta-Analyses of Disease Risk and Treatment Response. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 2555, pp.1–10.

North, C.M. & Christiani, D.C., 2014. Women and lung cancer: What's new? *Seminars in Thoracic and Cardiovascular Surgery*, 25(2), pp.1–13.

Ohlmeier, S. et al., 2012. Sputum proteomics identifies elevated PIGR levels in smokers and mild-to-moderate COPD. *Journal of Proteome Research*, 11(2), pp.599–608.

Pankov, R. & Kenneth, M., 2002. Fibronectin at a glance. *Journal of Cell Science*, 115, pp.3861–3863.

Papakonstantinou, E. & Karakiulakis, G., 2009. The "sweet" and "bitter" involvement of glycosaminoglycans in lung diseases : pharmacotherapeutic relevance. *British Journal of Pharmacology*, 157, pp.1111–1127.

Papi, A. et al., 2004. COPD increases the risk of squamous histological subtype in smokers who develop non-small cell lung carcinoma. *Thorax*, 59(8), pp.679–81.

Parameswaran, K. et al., 2006. Role of Extracellular Matrix and Its Regulators in Human Airway Smooth Muscle Biology. *Cell Biochemestry and Biophysics*, 44(7), pp.139–146.

Parra, E.R. et al., 2010. Association between decreases in type V collagen and apoptosis in mouse lung chemical carcinogenesis: a preliminary model to study cancer cell behavior. *Clinics*, 65(4), pp.425–432.

Paulissen, G. et al., 2009. Role of ADAM and ADAMTS metalloproteinases in airway diseases. *Respiratory research*, 10, p.127.

Pauwels, R.A. et al., 2012. NHLBI / WHO Workshop Summary Global Strategy for the Diagnosis , Management , and Prevention of Chronic Obstructive Pulmonary Disease NHLBI / WHO Global Initiative for Chronic Obstructive Lung Disease (GOLD) Workshop Summary. *American Journal of Respiratory and Critical Care Medicine*, 163, pp.1256–1276.

Pelosi, P. et al., 2007. The extracellular matrix of the lung and its role in edema formation. *Anais da Academia Brasileira de Ciências*, 79, pp.285–297.

Pierce, J.A. & Hocott, J.B., 1959. STUDIES ON THE COLLAGEN AND ELASTIN CONTENT OF THE HUMAN LUNG. *The Journal of Clinic Investigation*, 39(1), pp.8–14.

Plymoth, A. et al., 2006. Rapid proteome analysis of bronchoalveolar lavage samples of lifelong smokers and never-smokers by micro-scale liquid chromatography and mass spectrometry. *Clinical Chemistry*, 52(4), pp.671–679.

Rahman, I. & Adcock, I.M., 2006. Oxidative stress and redox regulation of lung inflammation in COPD. *European Respiratory Journal*, 28(1), pp.219–242.

Reiser, J., Adair, B. & Reinheckel, T., 2010. Specialized roles for cysteine cathepsins in health and disease. *The Journal of Clinical Investigation*, 120(10), pp.24–26.

Reunanen, N. & Kähäri, V., 2013. Matrix metalloproteinases in cancer cell invasion. Available at: https://www.ncbi.nlm.nih.gov/books/NBK6598/ [Accessed August 26, 2017].

Riaz, N. et al., 2016. Recurrent SERPINB3 and SERPINB4 mutations in patients who respond to anti-CTLA4 immunotherapy. *Nature genetics*, 48(11), pp.1327–1330.

Richards, S. et al., 2015. Standards and guidelines for the interpretation of sequence variants: a joint

consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17(5), pp.405–423.

Ridge, C., McErlean, A. & Ginsberg, M., 2013. Epidemiology of Lung Cancer. *Seminars in Interventional Radiology*, 30(2), pp.093–098.

Roberts, S.D. et al., 2006. FEV1/FVC ratio of 70% misclassifies patients with obstruction at the extremes of age. *Chest*, 130(1), pp.200–206.

Robinson, M.R., Wray, N.R. & Visscher, P.M., 2014. Explaining additional genetic variation in complex traits. *Trends in Genetics*, 30(4), pp.124–132.

Rocco, P.R.M. et al., 2001. Lung Tissue Mechanics and Extracellular Matrix Remodeling in Acute Lung Injury. , 164, pp.1067–1071.

Rock, K.L. & Kono, H., 2008. The inflammatory response to cell death. *Annual Review of Pathology*, 3, pp.99–126.

Rozario, T. & Desimone, D.W., 2011. The Extracellular Matrix In Development and Morphogenesis: A Dynamic View. *Developmental Biology*, 341(1), pp.126–140.

Schütz, A. et al., 2015. Lung adenocarcinomas and lung cancer cell lines show association of MMP-1 expression with STAT3 activation. *Translational Oncology*, 8(2), pp.97–105.

Schwarzbauer, J.E. & Desimone, D.W., 2011. Fibronectins, Their Fibrillogenesis, and In Vivo Functions. *Cold Spring Harbor Perspectives in Biology*, 3(7), pp.1–19.

Seixas, S., 2015. The Human SERPIN Repertoire and the Evolution of 14q32.1 and 18q21.3 Gene Clusters. In M. Geiger, F. Wahlmüller, & M. Furtmüller, eds. *The Serpin Family: Proteins with Multiple Functions in Health and Disease*. Cham: Springer International Publishing, pp. 1–14.

Shifren, A. & Mecham, R.P., 2006. The Stumbling Block in Lung Repair of Emphysema : Elastic Fiber Assembly. *Proccedings of the American Thoracic Society*, 3(5), pp.428-433.

Shintani, Y. et al., 2008. Collagen I promotes epithelial-to-mesenchymal transition in lung cancer cells via transforming growth factor-beta signaling. *American Journal of Respiratory Cell and Molecular Biology*, 38(1), pp.95–104.

Sinden, N.J. & Stockley, R.A., 2013. Proteinase 3 activity in sputum from subjects with alpha-1-antitrypsin deficiency and COPD. *European Respiratory Journal*, 41(5), pp.1042–1050.

Singh, S., Pillai, S. & Chellappan, S., 2011. Nicotinic Acetylcholine Receptor Signaling in Tumor Growth and Metastasis. *Journal of Oncology*, 2011, pp.1–11.

Smith, M.L. et al., 2007. Force-Induced Unfolding of Fibronectin in the Extracellular Matrix of Living Cells. *PLoS Biology*, 5(10), p.e268.

Soto-quiros, M.E. et al., 2009. MMP12, Lung Function, and COPD in High-Risk Populations. *The New England Journal of Medicine*, 361, pp.2599–2608.

Starcher, B.C., 2000. Activation of an Embryonic Gene Product in Pulmonary Emphysema * Identification of the Secreted. *Chest*, 117(5), p.229S–234S.

Starcher, B.C., Box, P.O. & Tx, T., 1986. Elastin and the lung. *Thorax*, 41, pp.577–585.

Stewart IV, J., 2001. Lung Carcinoma in African Americans: A Review of the Current Literature. *Cancer*, 91, pp.2476–2482.

Stratton, M.R., Campbell, P.J. & Andrew F, P., 2009. The cancer genome. *Nature*, 458(7239), pp.719–724.

Sun, Y., Sheshadri, N. & Zong, W.X., 2016. SERPINB3 and B4: From biochemistry to biology. *Seminars in Cell and Developmental Biology*, 62, pp.170-177.

Swinehart, I.T. & Badylak, S.F., 2017. Extracellular matrix bioscaffolds in tissue remodeling and morphogenesis. *Developmental Dynamics*, 245(3), pp.351–360.

Tallant, C., Marrero, A. & Gomis-Rüth, F.X., 2010. Matrix metalloproteinases: Fold and function of their catalytic domains. *Biochimica et Biophysica Acta - Molecular Cell Research*, 1803(1), pp.20–28.

The American Cancer Society, 2016. What Is Non-Small Cell Lung Cancer? Available at: https://www.cancer.org/cancer/non-small-cell-lung-cancer/about/what-is-non-small-cell-lung-cancer.html [Accessed August 24, 2017].

To, W.S. & Midwood, K.S., 2011. Plasma and cellular fibronectin : distinct and independent functions during tissue repair. *Fibrogenesis and Tissue Repair*, 4(21), pp.1–17.

Townsley, M.I., 2012. Structure and Composition of Pulmonary Arteries , Capillaries , and Veins. *Compreensive Physiology*, 2, pp.675–709.

Trupin, L. et al., 2003. The occupational burden of chronic obstructive pulmonary disease. *European Respiratory Journal*, 22(3), pp.462–469.

Turato, C. et al., 2009. Squamous cell carcinoma antigen-1 (SERPINB3) polymorphism in chronic liver disease. *Digestive and Liver Disease*, 41(3), pp.212–216.

Turato, C. & Pontisso, P., 2015. SERPINB3 (serpin peptidase inhibitor, clade B (ovalbumin), member 3). *Atlas of genetics and cytogenetics in oncology and haematology*, 19(3), pp.202–209.

U.S. Census Bureau, 2012. United States Census Bureau QuickFacts. , pp.1–9. Available at: https://www.census.gov/quickfacts/fact/table/US/PST045216 [Accessed September 23, 2017].

Vakkila, J. & Lotze, M.T., 2004. Opinion: Inflammation and necrosis promote tumour growth. *Nature Reviews Immunology*, 4(8), pp.641–648.

Vermaelen, K. & Brusselle, G., 2013. Exposing a deadly alliance: Novel insights into the biological links between COPD and lung cancer. *Pulmonary Pharmacology and Therapeutics*, 26(5), pp.544–554.

Vestbo, J. et al., 2013. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease GOLD executive summary. *American Journal of Respiratory and Critical Care Medicine*, 187(4), pp.347–365.

Vicary, G.W. et al., 2017. Nicotine stimulates collagen type I expression in lung via α7 nicotinic acetylcholine receptors. *Respiratory Research*, 18, pp.1–12.

Watson, W.H., Ritzenthaler, J.D. & Roman, J., 2016. Redox Biology Lung extracellular matrix and redox regulation. *Redox Biology*, 8, pp.305–315.

Wewers, M.D. & Crystal, R.G., 2013. Alpha-1 Antitrypsin Augmentation Therapy. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 10(sup1), pp.64–67.

Wolters, P.J. & Chapman, H. a, 2000. Importance of lysosomal cysteine proteases in lung disease. *Respiratory research*, 1(3), pp.170–177.

Yamamoto, S. et al, 1997. Immunohistochemical expression of SKALP/elafin in squamous cell carcinoma of the oesophagus. *British Journal of Cancer*, 76(8), pp.1081-1086

Yang, I.A., Holloway, J.W. & Fong, K.M., 2013. Genetic susceptibility to lung cancer and co-morbidities. *Journal of Thoracic Disease*, 5(SUPPL.5), pp.454-462.

Yoshida, N. et al., 2002. Immunohistochemical expression of SKALP/elafin in squamous cell carcinoma of human lung. *Oncology Reports*, 9(3), pp.495–501.

Yoshida, T. & Tuder, R., 2007. Pathobiology of cigarette smoke-induced chronic obstructive pulmonary disease. *Physiological reviews*, 87, pp.1047–1082.

Young, R.P. et al., 2015. Airflow limitation and histology shift in the National Lung Screening Trial: The NLST-ACRIN cohort substudy. *American Journal of Respiratory and Critical Care Medicine*, 192(9), pp.1060–1067.

Young, R.P. et al., 2011. Genetic evidence linking lung cancer and COPD: a new perspective. *The application of clinical genetics*, 4, pp.99–111.

Young, R.P., Hopkins, R.J. & Gamble, G.D., 2012. Clinical applications of gene-based risk prediction for lung cancer and the central role of chronic obstructive pulmonary disease. *Frontiers in Genetics*, 3(210), pp.1–7.

Ziółkowska-Suchanek, I. et al., 2015. Susceptibility loci in lung cancer and COPD: association of IREB2 and FAM13A with pulmonary diseases. *Scientific reports*, 5, p.13502.

5. Annexes

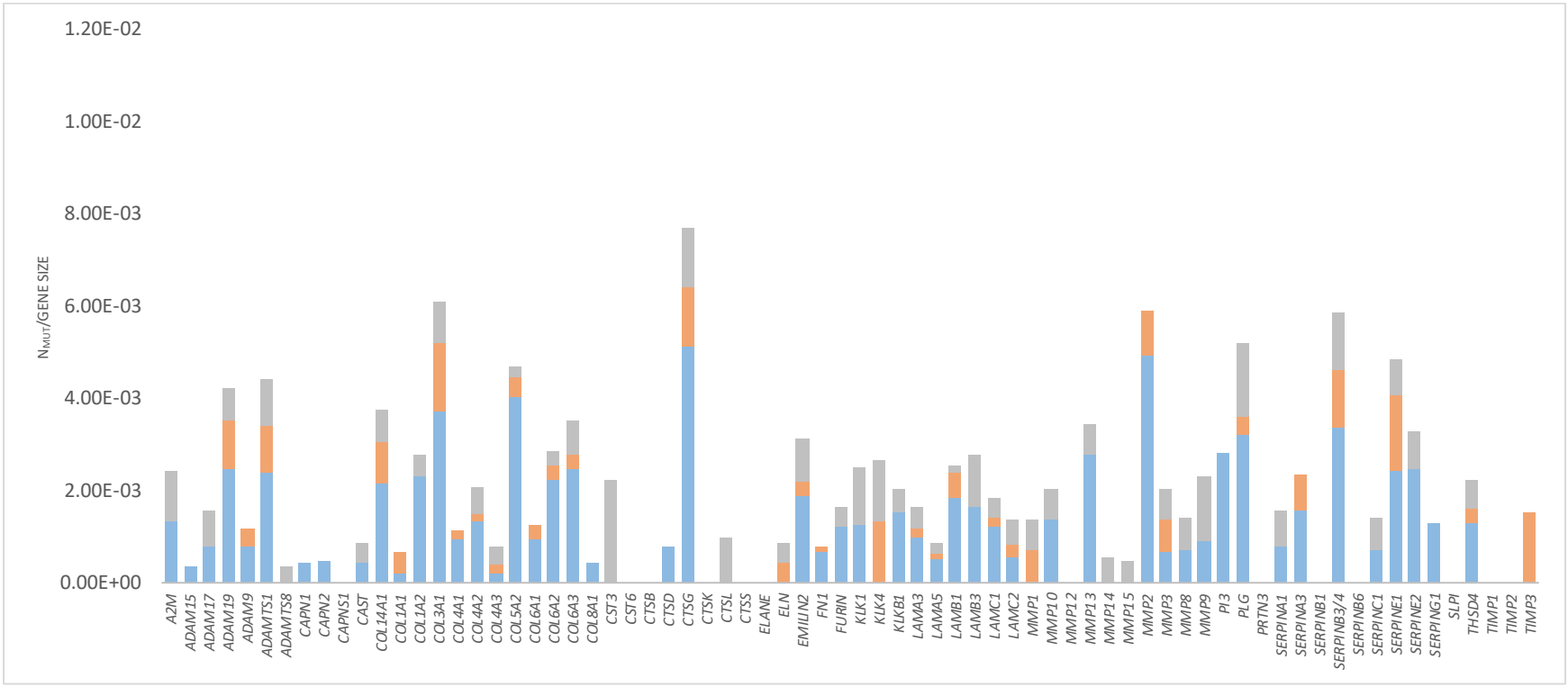

Figure A1. Somatic mutations rates of candidate genes in ADC patients with PolyPhen predictions. Mutation rates were normalized by gene size (coding sequencing and neighboring splice sites). Polyphen predictions as probably damaging mutations are shown in blue (●), as possibly damaging in orange (●), and benign in grey (●).

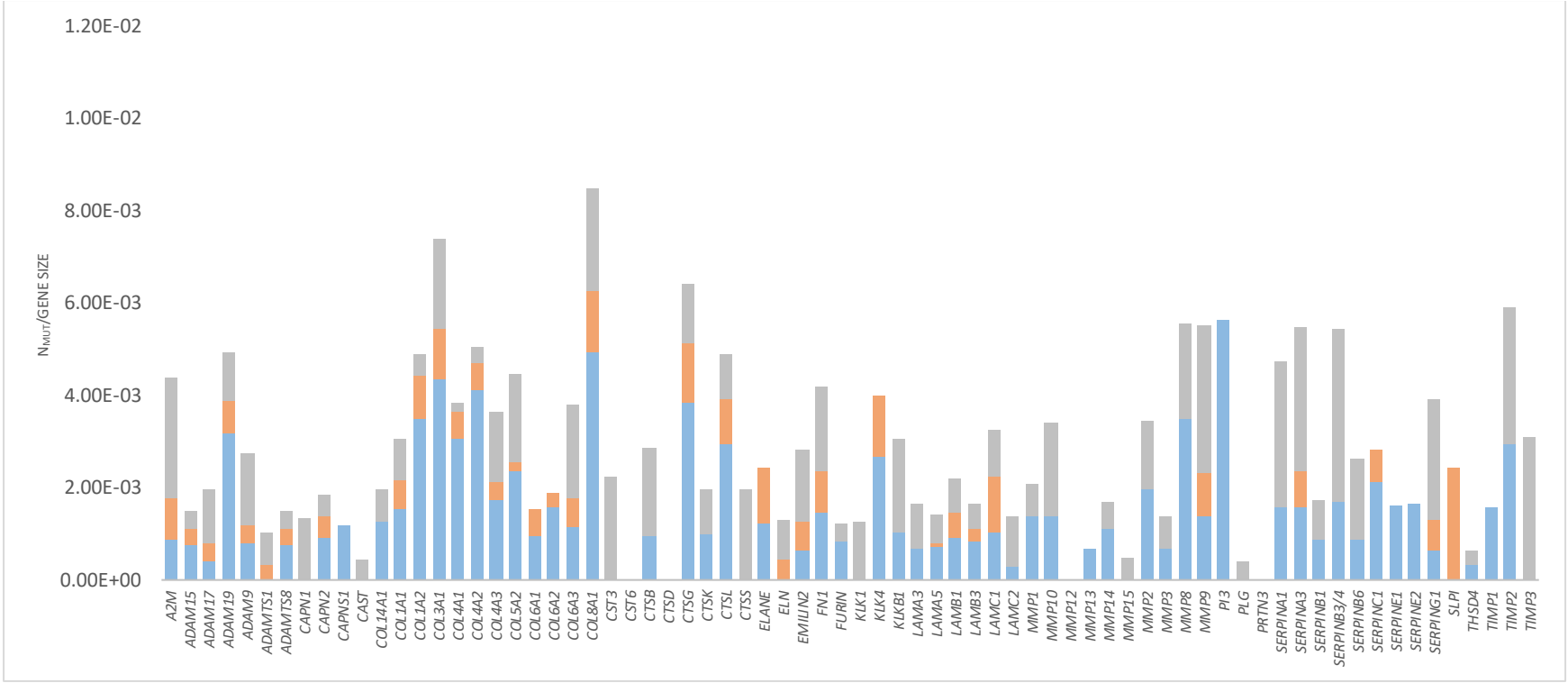

Figure A2. Somatic mutations rates of candidate genes in SCC patients with PolyPhen predictions. Mutation rates were normalized by gene size (coding sequencing and neighboring splice sites). Polyphen predictions as probably damaging mutations are shown in blue (●), as possibly damaging in orange (●), and benign in grey (●).

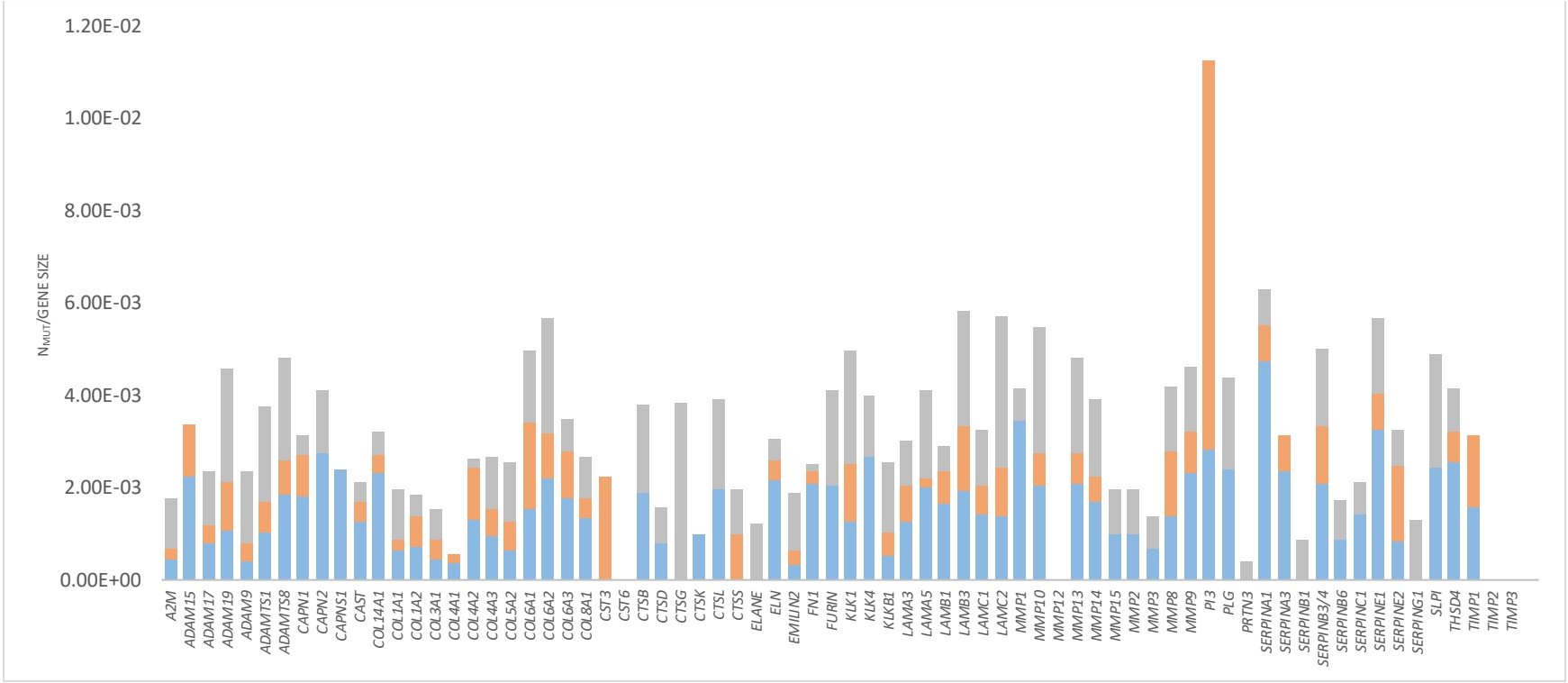

Figure A3. Germline mutations rates of candidate genes in ADC patients with PolyPhen predictions. Mutation rates were normalized by gene size (coding sequencing and neighboring splice sites). A cutoff of MAF <0.05 of non-Finnish European population from ExAc, excluding common variants was used. Polyphen predictions as probably damaging mutations are shown in blue (●), as possibly damaging in orange (●), and benign in grey (●).

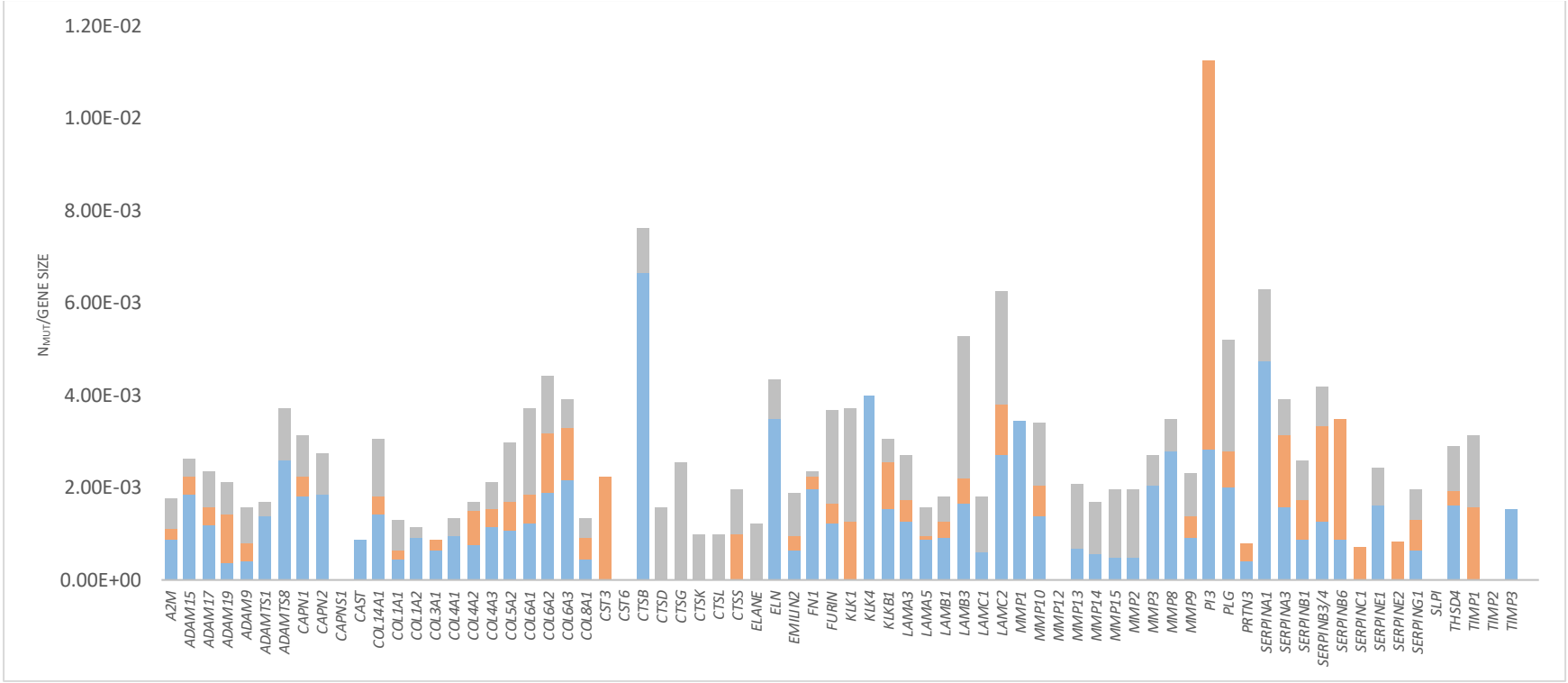

Figure A4. Germline mutations rates of candidate genes in SCC patients with PolyPhen predictions. Mutation rates were normalized by gene size (coding sequencing and neighboring splice sites). A cutoff of MAF <0.05 of non-Finnish European population from ExAc, excluding common variants was used. Polyphen predictions as probably damaging mutations are shown in blue (●), as possibly damaging in orange (●), and benign in grey (●).

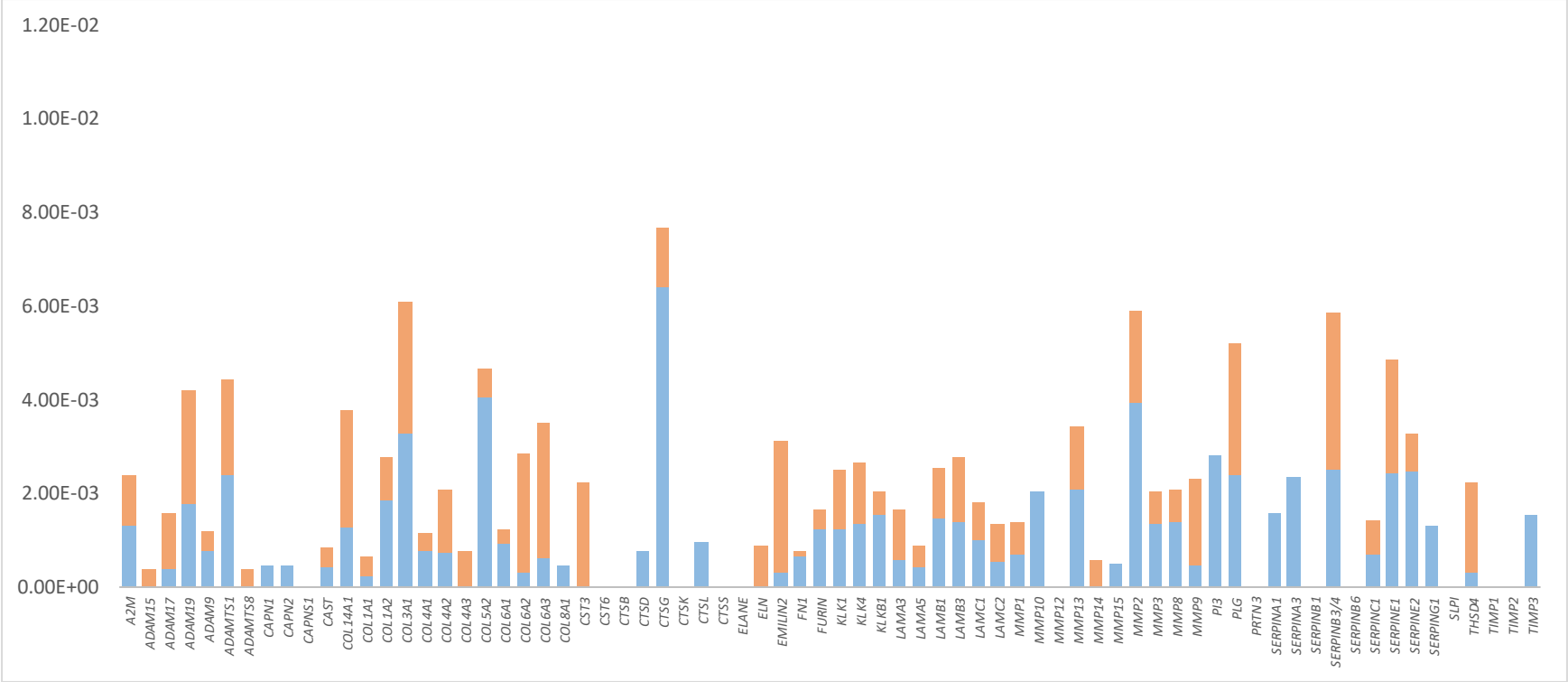

Figure A5. Somatic mutations rates of candidate genes in ADC patients with SIFT predictions. Mutation rates were normalized by gene size (coding sequencing and neighboring splice sites). SIFT predictions as deleterious mutations are shown in blue (●), and as tolerated in orange (●).

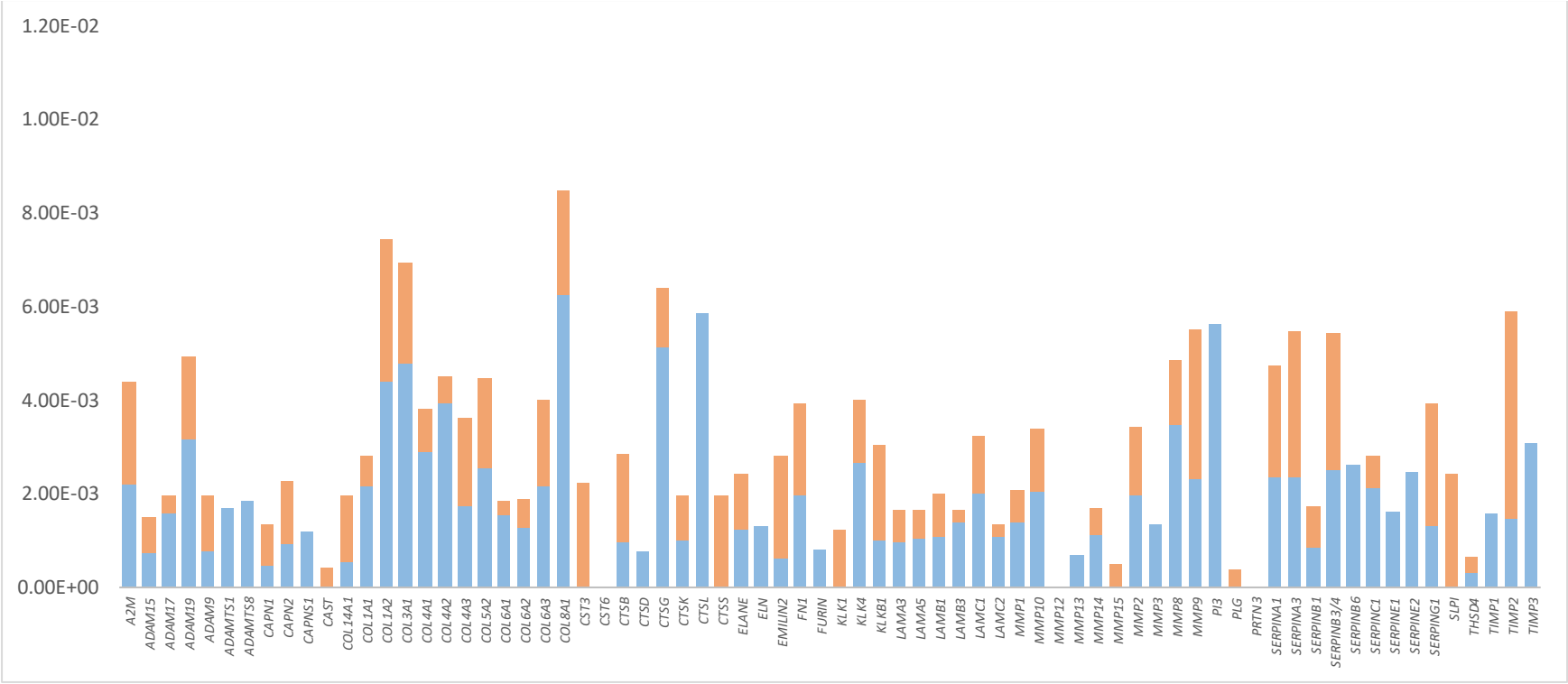

Figure A6. Somatic mutations rates of candidate genes in SCC patients with SIFT predictions. Mutation rates were normalized by gene size (coding sequencing and neighboring splice sites). SIFT predictions as deleterious mutations are shown in blue (●), and as tolerated in orange (●).

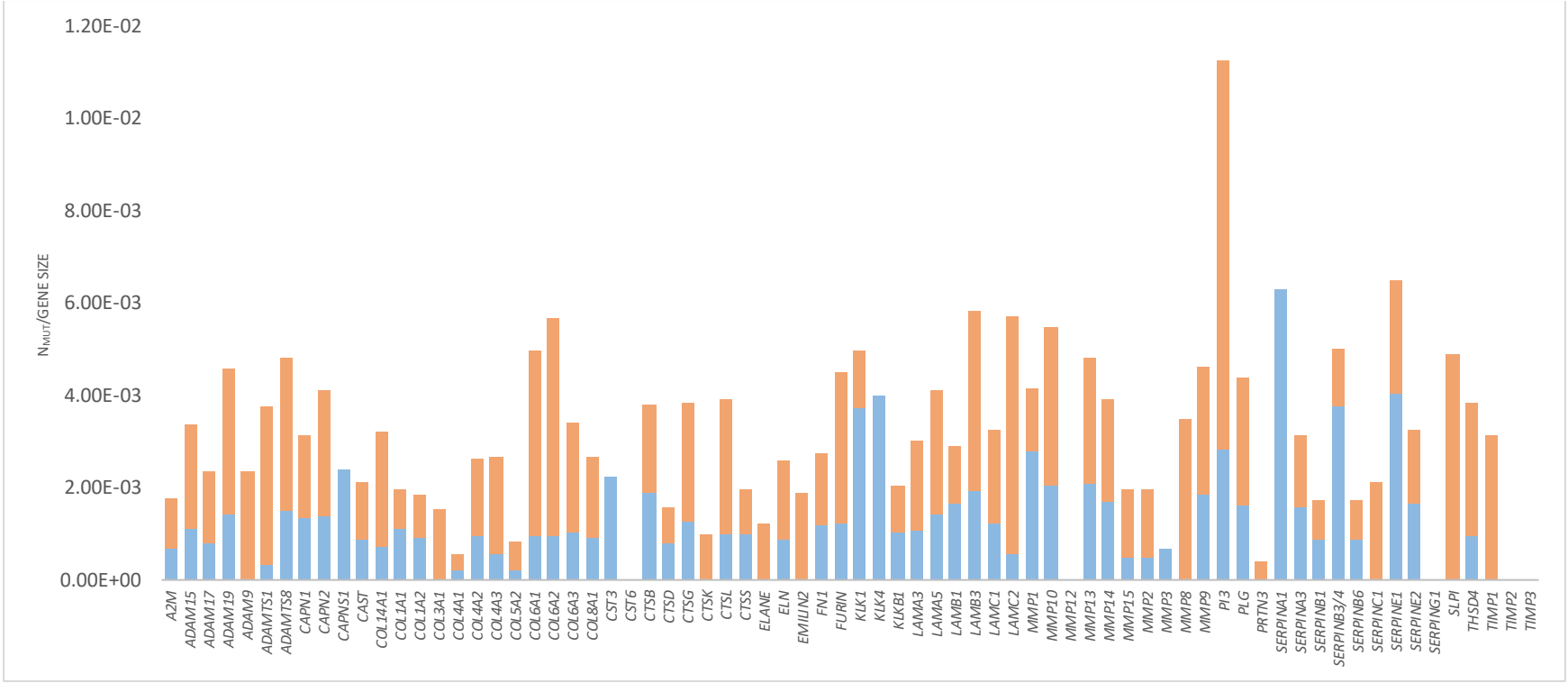

Figure A7. Germline mutations rates of candidate genes in ADC patients with SIFT predictions. Mutation rates were normalized by gene size (coding sequencing and neighboring splice sites). A cutoff of MAF <0.05 of non-Finnish European population from ExAc, excluding common variants was used. SIFT predictions as deleterious mutations are shown in blue (●), and as tolerated in orange (●).

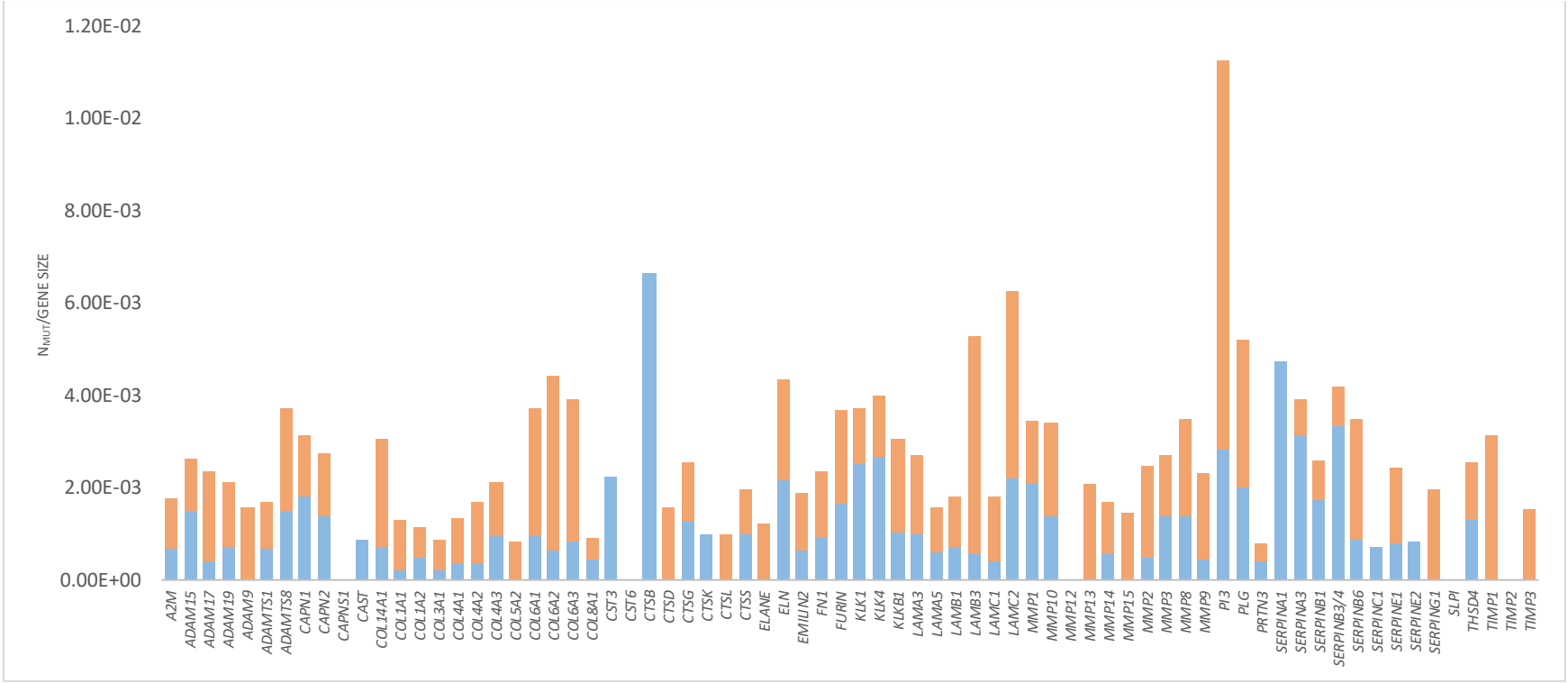

Figure A8. Germline mutations rates of candidate genes in SCC patients with SIFT predictions. Mutation rates were normalized by gene size (coding sequencing and neighboring splice sites). A cutoff of MAF <0.05 of non-Finnish European population from ExAc, excluding common variants was used. SIFT predictions as deleterious mutations are shown in blue (●), and as tolerated in orange (●).

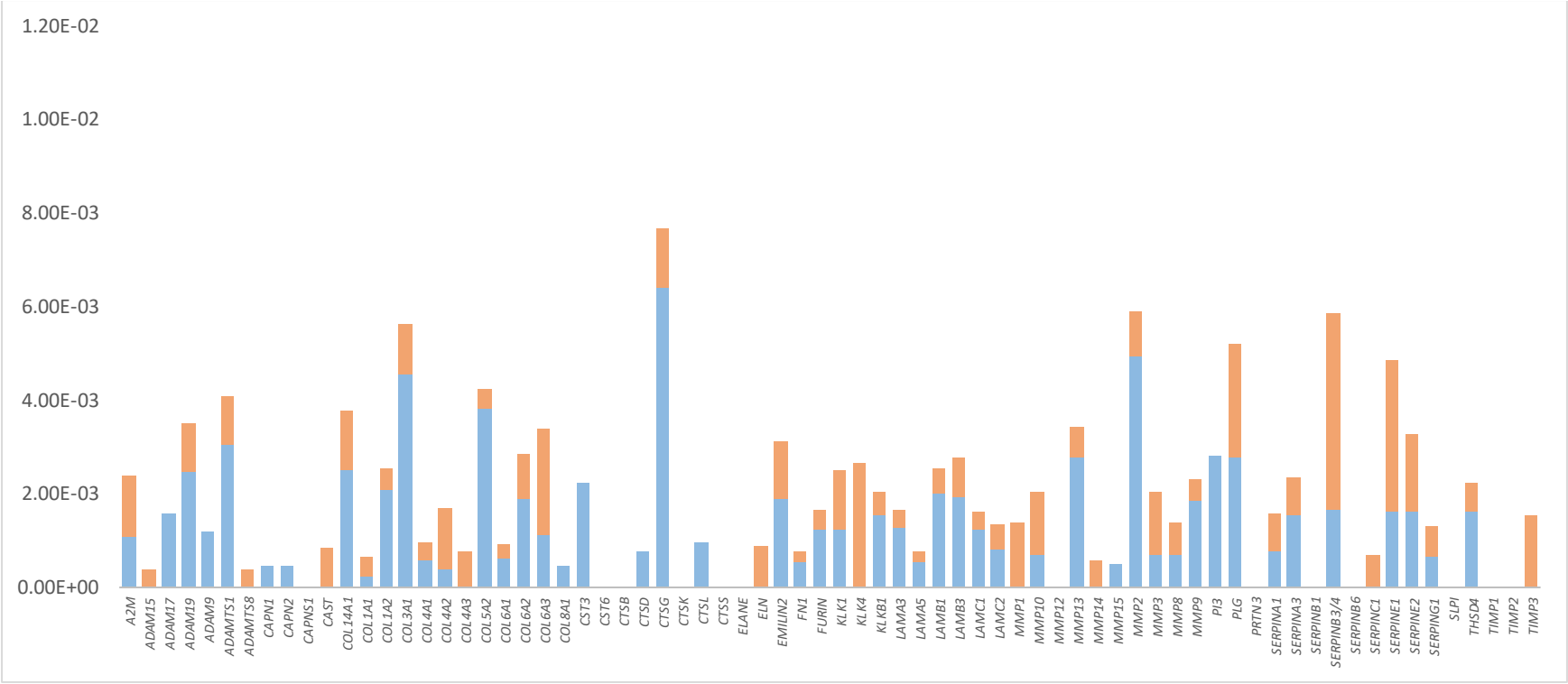

Figure A9. Somatic mutations rates of candidate genes in ADC patients with CADD predictions. Mutation rates were normalized by gene size (coding sequencing and neighboring splice sites). CADD PHRED-like scores equal or above 14.5, considered as deleterious mutations are shown in blue (●), and scores bellow 14.5, considered as tolerated in orange (●).

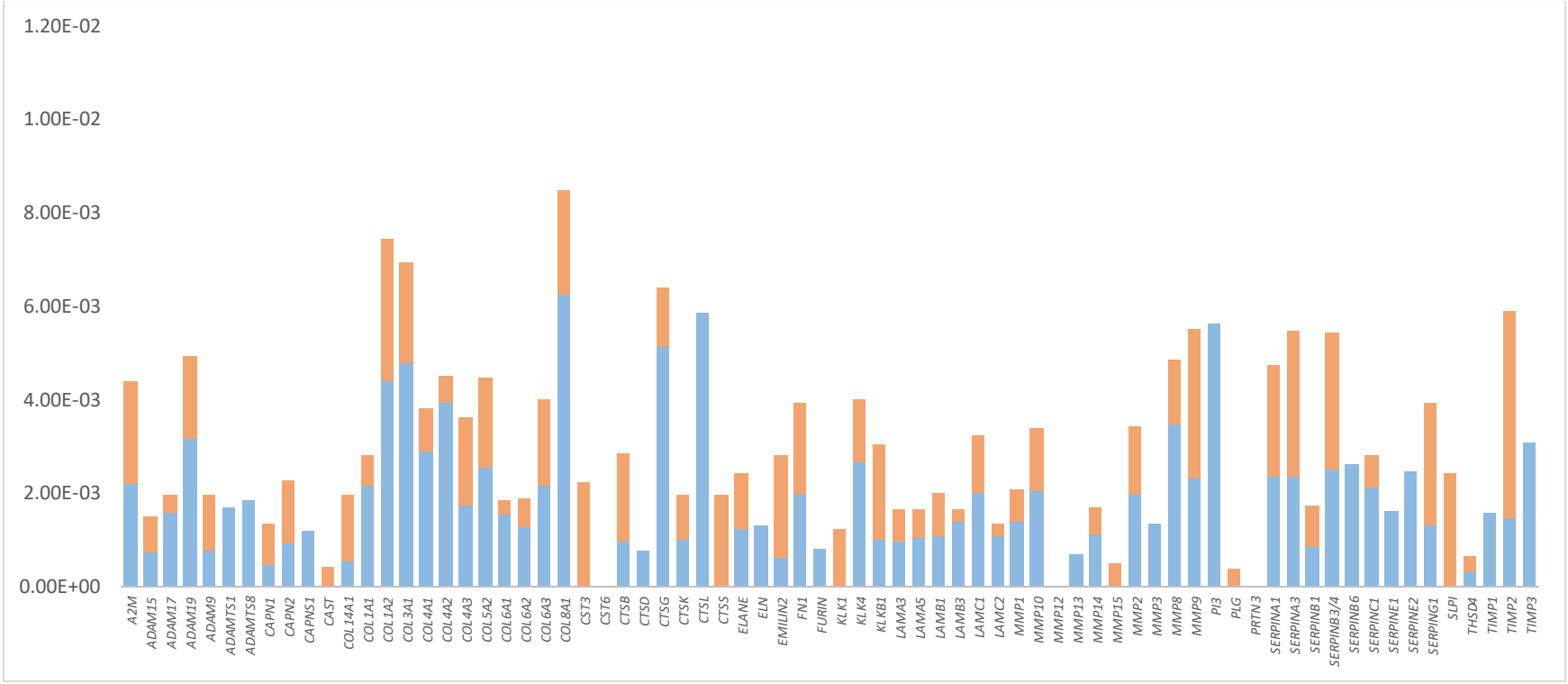

Figure A10. Somatic mutations rates of candidate genes in SCC patients with CADD predictions. Mutation rates were normalized by gene size (coding sequencing and neighboring splice sites). SIFT predictions as deleterious mutations are shown in blue (●), and as tolerated in orange (●).

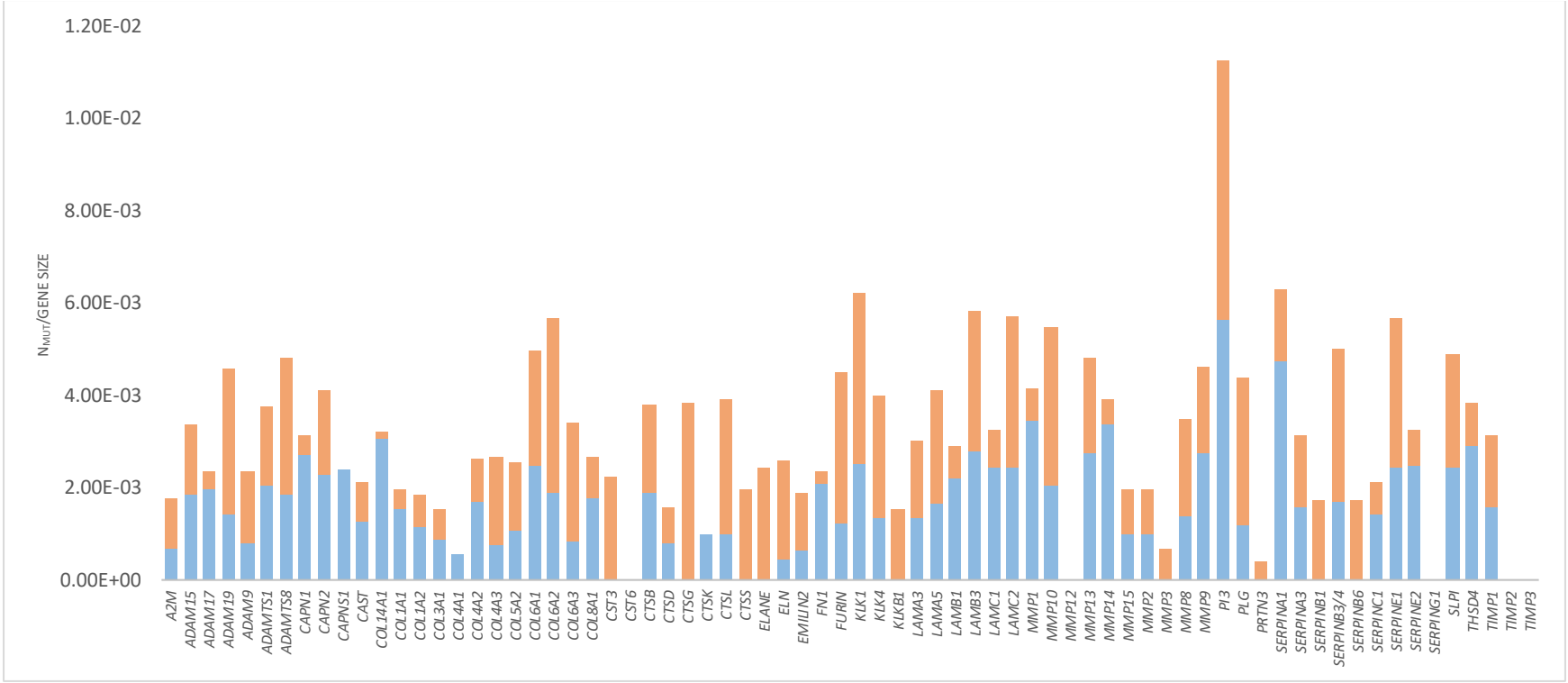

Figure A11. Germline mutations rates of candidate genes in ADC patients with CADD predictions. Mutation rates were normalized by gene size (coding sequencing and neighboring splice sites). A cutoff of MAF <0.05 of non-Finnish European population from ExAc, excluding common variants was used. CADD PHRED-like scores equal or above 14.5, considered as deleterious mutations are shown in blue (●), and scores bellow 14.5, considered as tolerated in orange (●).

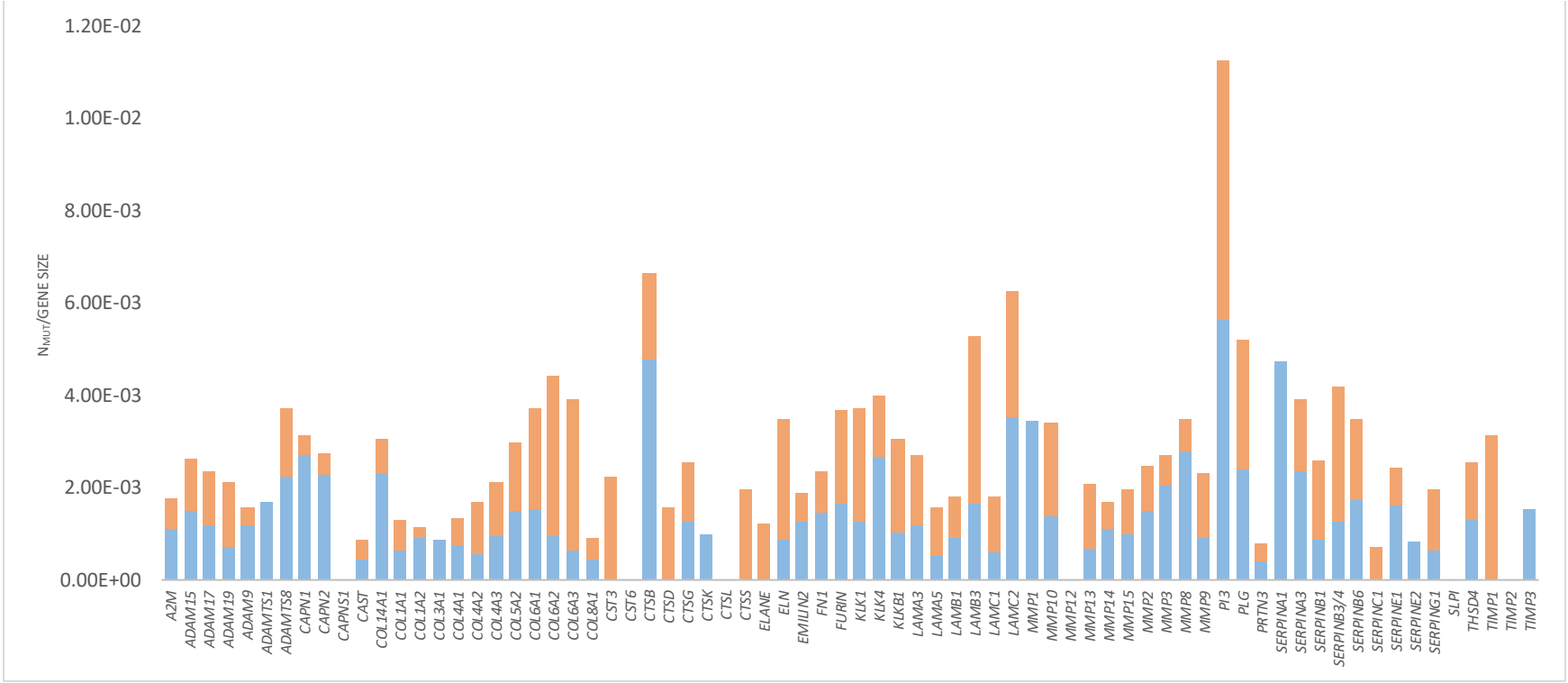

Figure A12. Germline mutations rates of candidate genes in SCC patients with CADD predictions. Mutation rates were normalized by gene size (coding sequencing and neighboring splice sites). A cutoff of MAF <0.05 of non-Finnish European population from ExAc, excluding common variants was used. CADD PHRED-like scores equal or above 14.5, considered as deleterious mutations are shown in blue (●), and scores bellow 14.5, considered as tolerated in orange (●).

Table T1. PCR conditions for SERPINB3/B4 amplification.

Gene(s)	Fragments	PCR Conditions
SERPINB3/B4	**A**	95ºC 5min; 35 cycles: 94ºC 30s, 55ºC 30s, 68ºC 2min30s; 68ºC 20min.
	B	95ºC 5min; 35 cycles: 94ºC 30s, 55ºC 30s, 68ºC 1min30s; 68ºC 20min.
	C	95ºC 5min; 35 cycles: 94ºC 30s, 52ºC 30s, 68ºC 45s; 68ºC 20min.
	D	95ºC 5min; 35 cycles: 94ºC 30s, 52ºC 30s, 68ºC 45s; 68ºC 20min.
SERPINB3	**E**	95ºC 5min; 35 cycles: 94ºC 30s, 55ºC 30s, 68ºC 1min30s; 68ºC 20min.
SERPINB4	**E**	95ºC 5min; 40 cycles: 94ºC 30s, 55ºC 30s, 68ºC 1min30s; 68ºC 20min.

Table T2. Semi-nested PCR conditions for SERPINB3/B4 amplification.

Gene(s)	Fragments	PCR Conditions
SERPINB3/B4	**A**	95ºC 5min; 35 cycles: 94ºC 30s, 54ºC 30s, 68ºC 2min30s; 68ºC 20min.
	B	95ºC 5min; 35 cycles: 94ºC 30s, 54ºC 30s, 68ºC 1min30s; 68ºC 20min.
	E1	95ºC 5min; 35 cycles: 94ºC 30s, 54ºC 30s, 68ºC 1min30s; 68ºC 20min.
	E2	95ºC 5min; 35 cycles: 94ºC 30s, 54ºC 30s, 68ºC 1min30s; 68ºC 20min.

Table T3. Sequencing PCR conditions for the three gene amplification.

Genes	PCR Conditions
SERPINB3, SERPINB4* and *CTSG	6 cycles: 96ºC 10s, 65ºc 1min; 6 cycles: 96ºC 10s, 64ºc 1min; 6 cycles: 96ºC 10s, 63ºc 1min; 6 cycles: 96ºC 10s, 62ºc 1min; 6 cycles: 96ºC 10s, 61ºc 1min; 6 cycles: 96ºC 10s, 60ºc 1min; 6 cycles: 96ºC 10s, 59ºc 1min; 6 cycles: 96ºC 10s, 58ºc 1min; 6 cycles: 96ºC 10s, 57ºc 1min; 6 cycles: 96ºC 10s, 56ºc 1min; 6 cycles: 96ºC 10s, 55ºc 1min.